

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2001年11月30日

出 願 番 号

Application Number:

特願2001-368002

[ ST.10/C ]:

[ JP2001-368002 ]

出 願 人

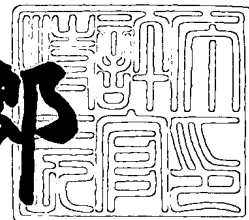
Applicant(s):

大森 聡

2003年 3月14日

特許庁長官  
Commissioner,  
Japan Patent Office

太田信一郎



出証番号 出証特2003-3017023

【書類名】 特許願

【整理番号】 2001A16

【提出日】 平成13年11月30日

【あて先】 特許庁長官殿

【発明者】

    【住所又は居所】 埼玉県さいたま市西堀4丁目11番7号627

    【氏名】 大森 聡

【特許出願人】

    【識別番号】 300000513

    【氏名又は名称】 大森 聡

【パリ条約による優先権等の主張】

    【国名】 欧州特許庁

    【出願日】 2001年 4月18日

    【出願番号】 PCT/JP01/03324

【手数料の表示】

    【予納台帳番号】 095958

    【納付金額】 21,000円

【提出物件の目録】

    【物件名】 明細書 1

    【物件名】 図面 1

    【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 生物学的物質の配列情報の記録方法及び装置、前記配列情報の供給方法、並びに前記配列情報を記録した記録媒体

【特許請求の範囲】

【請求項 1】 生物学的物質の配列情報の記録方法であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが  $m$  ビット ( $m$  は 16 以上の整数) の部分データに分割し、

複数行の前記部分データに各行毎に前記非配列方向にガロア体  $GF(2^m)$  上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、

複数列の前記部分データに各列毎に前記配列方向にガロア体  $GF(2^m)$  上の第 2 の演算を施して第 2 組のパリティ情報を求め、

前記第 1 組及び第 2 組のパリティ情報で前記生物学的物質の配列を表すことを特徴とする生物学的物質の配列情報の記録方法。

【請求項 2】 請求項 1 に記載の記録方法であって、

前記ガロア体  $GF(2^m)$  上の生成元を  $\alpha$  としたとき、

前記第 1 組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次  $\alpha^{sp}$ ,  $\alpha^{s(p+1)}$ ,  $\alpha^{s(p+2)}$ , ...,  $\alpha^{s(p+dp)}$  ( $s$  は 0 以上の整数、 $p$  は 0 以上の整数、 $dp$  は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

前記第 2 組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次  $\alpha^{tq}$ ,  $\alpha^{t(q+1)}$ ,  $\alpha^{t(q+2)}$ , ...,  $\alpha^{t(q+dq)}$  ( $t$  は 0 以上の整数、 $q$  は 0 以上の整数、 $dq$  は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の記録方法。

【請求項 3】 請求項 2 に記載の記録方法であって、

前記整数  $s$  及び  $t$  は 0 であることを特徴とする生物学的物質の配列情報の記録方法。

【請求項 4】 請求項 2 に記載の記録方法であって、  
前記整数  $s$  及び  $t$  は 1 であることを特徴とする生物学的物質の配列情報の記録方法。

【請求項 5】 請求項 2 に記載の記録方法であって、  
前記第 1 組のパリティ情報は、前記複数行の各行毎に前記整数  $s$  について互いに異なる複数の値で求めた複数の和を含み、

前記第 2 組のパリティ情報は、前記複数列の各列毎に前記整数  $t$  について互いに異なる複数の値で求めた複数の和を含むことを特徴とする生物学的物質の配列情報の記録方法。

【請求項 6】 請求項 1 ～ 5 の何れか一項に記載の記録方法であって、  
前記生物学的物質の配列中の各生物学的物質をそれぞれ 6 ビット以下の数値データで表して得られる数値データを前記演算の対象とすることを特徴とする生物学的物質の配列情報の記録方法。

【請求項 7】 請求項 1 ～ 6 の何れか一項に記載の記録方法であって、  
前記ガロア体  $GF(2^m)$  を規定する整数  $m$  は 64 の倍数であることを特徴とする生物学的物質の配列情報の記録方法。

【請求項 8】 請求項 1 ～ 7 の何れか一項に記載の記録方法であって、  
前記生物学的物質の配列を基準配列として、該基準配列の前記 2 組のパリティ情報に対応させて、検査対象の生物学的物質の配列について前記 2 組のパリティ情報を求め、

前記 4 組のパリティ情報より前記基準配列に対する前記検査対象の生物学的物質の配列の相違部を求めることを特徴とする生物学的物質の配列情報の記録方法。

【請求項 9】 請求項 1 ～ 8 の何れか一項に記載の記録方法であって、  
前記生物学的物質は、DNA、RNA、又は遺伝子の少なくとも一部を構成するヌクレオチドであることを特徴とする生物学的物質の配列情報の記録方法。

【請求項 10】 請求項 1 ～ 8 の何れか一項に記載の記録方法であって、  
前記生物学的物質は、一つのタンパク質の少なくとも一部を構成するアミノ酸であることを特徴とする生物学的物質の配列情報の記録方法。

【請求項 1 1】 生物学的物質の配列情報の記録装置であって、

前記生物学的物質の配列情報を読み取る配列読み取り装置と、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが  $m$  ビット ( $m$  は 16 以上の整数) の部分データに分割するデータ配列手段と、

複数行の前記部分データに各行毎に前記非配列方向にガロア体  $GF(2^m)$  上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体  $GF(2^m)$  上の第 2 の演算を施して第 2 組のパリティ情報を求める演算手段と、

前記第 1 組及び第 2 組のパリティ情報を記録媒体に記録する記録手段とを有することを特徴とする生物学的物質の配列情報の記録装置。

【請求項 1 2】 請求項 1 1 に記載の記録装置であって、

前記ガロア体  $GF(2^m)$  上の生成元を  $\alpha$  としたとき、

前記第 1 組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次  $\alpha^{sp}$ ,  $\alpha^{s(p+1)}$ ,  $\alpha^{s(p+2)}$ , ...,  $\alpha^{s(p+dp)}$  ( $s$  は 0 以上の整数、 $p$  は 0 以上の整数、 $dp$  は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

前記第 2 組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次  $\alpha^{tq}$ ,  $\alpha^{t(q+1)}$ ,  $\alpha^{t(q+2)}$ , ...,  $\alpha^{t(q+dq)}$  ( $t$  は 0 以上の整数、 $q$  は 0 以上の整数、 $dq$  は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の記録装置。

【請求項 1 3】 生物学的物質の配列情報を記録したコンピュータ読み取り可能な記録媒体であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが  $m$  ビット ( $m$  は 16 以上の整数) の部分データに分割し、

複数行の前記部分データに各行毎に前記非配列方向にガロア体  $GF(2^m)$  上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体  $GF(2^m)$  上の第 2 の演算を施して第 2 組のパリティ情報を求め、

前記生物学的物質の配列に関する情報が、前記第 1 組及び第 2 組のパリティ情報として記録されたことを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項 1 4】 請求項 1 3 に記載の記録媒体であって、

前記生物学的物質の配列に対応する前記テキストデータ、又は該テキストデータに対応する前記数値データの 4 0 ビット以上の長さの数学的な要約値が更に前記記録媒体に記録されたことを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項 1 5】 生物学的物質の配列情報の供給方法であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを保持する供給者が、前記テキストデータ、又はこれに対応する前記数値データを第 1 ファイルに記録して保持する第 1 ステップと、

前記第 1 ファイルに記録されている前記テキストデータ、又は該テキストデータに対応する前記数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが  $m$  ビット ( $m$  は 1 6 以上の整数) の部分データに分割し、

複数行の前記部分データに各行毎に前記非配列方向にガロア体  $GF(2^m)$  上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体  $GF(2^m)$  上の第 2 の演算を施して第 2 組のパリティ情報を求める第 2 ステップと、

前記供給者が、前記第 1 組及び第 2 組のパリティ情報を第 2 ファイルに記録して保持する第 3 ステップと、

前記生物学的物質の配列情報のユーザが、通信回線を介して前記供給者より前記第 2 ファイルに記録されている前記 2 組のパリティ情報を受け取る第 4 ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

【請求項 1 6】 請求項 1 5 に記載の供給方法であって、

前記ユーザが、前記 2 組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内の前記供給者の生物学的物質の配列情報との相違部を特定する第 5 ステップと、

該相違部の配列の復元ができない場合に、前記ユーザが前記通信回線を介して前記供給者より前記第 1 ファイルに記録されている前記テキストデータ、又は前記数値データの内の前記配列の復元ができない部分の配列情報を受け取る第 6 ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

【請求項 1 7】 請求項 1 5 又は 1 6 に記載の供給方法であって、

前記ガロア体  $GF(2^m)$  上の生成元を  $\alpha$  としたとき、

前記第 1 組のパリティ情報は、複数列の各行の前記部分データにそれぞれ前記非配列方向に順次  $\alpha^{sp}$ ,  $\alpha^{s(p+1)}$ ,  $\alpha^{s(p+2)}$ , ...,  $\alpha^{s(p+dp)}$  ( $s$  は 0 以上の整数、 $p$  は 0 以上の整数、 $d_p$  は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

前記第 2 組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次  $\alpha^{tq}$ ,  $\alpha^{t(q+1)}$ ,  $\alpha^{t(q+2)}$ , ...,  $\alpha^{t(q+dq)}$  ( $t$  は 0 以上の整数、 $q$  は 0 以上の整数、 $d_q$  は 1 以上の整数) を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の供給方法。

【請求項 1 8】 請求項 1 5、1 6、又は 1 7 に記載の供給方法であって、

前記供給者は、前記生物学的物質の配列の長さの情報、及び前記配列を表すテキストデータ又は前記数値データの数学的な要約値の情報を前記通信回線を介して閲覧可能な状態にしておき、

前記ユーザは、前記第 4 ステップの前に前記通信回線を介して前記配列の長さの情報及び前記数学的な要約値の情報を閲覧することを特徴とする生物学的物質の配列情報の供給方法。

【発明の詳細な説明】

## 【 0 0 0 1 】

## 【発明の属する技術分野】

本発明は、例えばDNA（デオキシリボ核酸：deoxyribonucleic acid）、RNA（リボ核酸：ribonucleic acid）、若しくは遺伝子等の核酸の少なくとも一部を構成する一列のヌクレオチド又はタンパク質の少なくとも一部を構成する一列のアミノ酸などの生物学的物質（Biological materials）の配列情報の記録方法及び装置に関する。更に本発明は、その配列情報を供給するためのビジネスモデルとして好適な配列情報の供給方法、及びその配列情報を記録するコンピュータ読み取り可能な記録媒体に関する。

## 【 0 0 0 2 】

## 【従来の技術】

人間、及び他の生物（動物、植物、微生物等）のDNAを構成する1対のヌクレオチドの鎖（又は塩基の鎖）の配列情報の解読が世界的に行われている。この場合、従来よりDNAを構成する4種類のヌクレオチドは、塩基としてアデニンを含むヌクレオチド、グアニンを含むヌクレオチド、シトシンを含むヌクレオチド、及びチミンを含むヌクレオチドにそれぞれ文字A、G、C、及びTを割り当てることによって、それぞれ1バイト（＝8ビット）のテキストデータで表わされている。その結果として一つのDNAの配列は、それを構成する1対の重合体の鎖の内的一方の鎖のヌクレオチド（n個とする）の配列を順次文字A、G、C、T（又はa、g、c、t）の何れかで表すことによって、nバイトのテキストデータで表されていた。同様に、一つのRNAを構成する1本のn個のヌクレオチドの配列は、チミンを含むヌクレオチドの代わりにウラシルを含むヌクレオチドに文字U（又はu）を割り当てることによって、nバイトのテキストデータで表されていた。

## 【 0 0 0 3 】

これに関して、例えば人間の最も大きい第1染色体中のDNAの配列は、約2億5千万個のヌクレオチドの配列であり、最も小さい第22染色体中のDNAの配列は、約5000万個のヌクレオチドの配列であるため、人間の各染色体中のDNAの配列は、約250Mバイト～50Mバイトのテキストデータで表すこと



ができる。更に、一人の人間の全部のDNA情報（ゲノム）は、約30億個のヌクレオチドの配列で表すことができるため、そのゲノムは、約3Gバイトのテキストデータで記録することができる。なお、それらのテキストデータに対して通常のファイル圧縮技術を適用することによって、それらのテキストデータは、例えば元のデータの50%程度の圧縮ファイルとしても記録、又は送信することができる。

#### 【0004】

また、DNAの配列の解読に続いて、DNA中の多数の遺伝子の情報に基づいてそれぞれ合成されるタンパク質の機能の研究も広く行われている。この場合、タンパク質を構成する20種類のアミノ酸は、三文字表記（3-Letter Code）ではそれぞれ3文字（例えばAla, Cys, Glu等）のテキストデータで表され、一文字表記（1-Letter Code）ではそれぞれ1文字のテキストデータ（例えばA, C, E等）で表されるため、n個のアミノ酸よりなるタンパク質の配列は、nバイトのテキストデータで表すことができる。そして、種々のタンパク質は、それらのアミノ酸が約20個～約1000個程度所定の順序で配列されたものであるため、それらのタンパク質の配列は、最大でも約1kバイト程度のテキストデータで記録することができる。また、例えば人間の遺伝子の総数は約3万個と言われており、タンパク質は理論的なものも含めて約10万種類の存在が可能であると言われている。

#### 【0005】

##### 【発明が解決しようとする課題】

上記の如く例えば一人の人間のDNA情報をテキストデータで記録するためには、全部で3Gバイト程度の記憶容量が必要であり、仮に通常の圧縮ファイルの技術を適用しても1Gバイト程度の記憶容量が必要である。また、人間以外の大腸菌や各種ウイルス等のDNA情報も解析されて次第に公開されるようになっていくが、これらのDNA情報をテキストデータの形で多く集めると、数100Gバイト程度の記憶容量が必要である。これはRNAの配列情報についても同様である。

#### 【0006】

このように人間又は他の生物のDNA情報をテキストデータ、又はこの通常の圧縮ファイルの形で記録するものとする、例えば1枚の記憶容量が5Gバイト程度のDVD-ROM(digital video disc-ROM)ディスクのように膨大な記憶容量を持つ記録媒体が必要である。更に、そのDNA情報を利用する場合にその記録媒体からの読み出し時間が長くなり、処理時間が長くなるという不都合がある。

【0007】

また、現状の一般の通信回線の通信速度は、最大で5Mbps程度であるため、例えば1Gバイト程度のDNA情報をその通信回線を介して送信するものとする、送信時間は最短でも約30分程度となる。特に最近はそのDNA情報をデジタルの携帯電話システムを介して送信する場合も考えられるが、現在の携帯電話システムの通信速度はせいぜい1Mbps程度であるため、少なくとも人間のDNA情報の伝送で使用することは現状ではあまり実用的ではない。

【0008】

次に、例えば或る微生物のDNA中の遺伝子について複数の研究者が並行して研究するような場合に、複数の研究者が保有している標準となるDNAのヌクレオチドの配列の同一性をどのように保証するのかという問題がある。即ち、そのDNAのヌクレオチドの配列が例えば数Mバイト（文字数で数100万文字）程度のテキストデータで記録されている場合に、複数の研究者が互いに自分のテキストデータと他人のテキストデータとの同一性（完全一致性）を短時間に確認するのは必ずしも容易ではない。

【0009】

これに関連して、例えば人間又は他の生物のDNA情報の利用方法としては、標準的なDNAの配列と、検査対象のDNAの配列との間の相違する部分をサーチする場合が考えられる。これは、いわゆるSNP（一塩基変位多型：Single Nucleotide Polymorphism）の可能性を検査するような場合に必要になると考えられる。しかしながら、両方のDNAのヌクレオチドの配列がそれぞれ膨大なテキストデータで表わされている場合に、それら2つのテキストデータを比較して相違点を検出するにはかなりの長い時間が必要となり、検査時間が長くなるという

不都合がある。

【0010】

更に、人間又は他の生物のDNA情報を製薬会社の研究者等のユーザに提供するビジネスも行われつつあるが、この場合に、情報供給者が例えば通信回線を通じてDNA情報をユーザに提供する場合には、できるだけ少ない情報量で、即ち短い送信時間で必要な情報をユーザに提供できるビジネスモデルが必要である。また、ユーザ側では、提供されたDNA情報に伝送エラー等が無いかどうかを容易に確認できることが望ましい。上記の各課題はRNAのヌクレオチドの配列情報についても同様に当てはまるものである。

【0011】

更に、一つのタンパク質のアミノ酸の配列は、最大でも約1kバイト程度のテキストデータで記録することができるが、タンパク質の種類は理論的に約10万个程度にもなるため、全部のタンパク質の配列情報をテキストデータで表すと、全部のDNAの配列情報程度の膨大な量となる。従って、個々のタンパク質の配列は、できるだけ少ない情報量で記録できることが望ましい。また、2つのタンパク質の配列情報の相違部を容易に確認できるシステムも必要である。

【0012】

本発明は斯かる点に鑑み、核酸中の一列のヌクレオチド、又はタンパク質中の一列のアミノ酸などの生物学的物質の配列情報を近似的に少ないデータ量で記録できる記録方法及び装置を提供することを第1の目的とする。

また、本発明は、2つの生物学的物質の配列情報同士の相違する部分を少ないデータ量で容易に検出できると共に、必要に応じてその相違する部分の情報を復元できる記録方法及び装置を提供することを第2の目的とする。

【0013】

また、本発明は、一列のヌクレオチド、又は一列のアミノ酸などの生物学的物質の配列情報をユーザに提供する場合に、ユーザが提供された配列情報と情報供給者が保持している配列情報との相違する部分を少ないデータ量で容易に確認できるビジネスモデル（情報供給方法）を提供することを第3の目的とする。

また、本発明は、生物学的物質の配列情報が少ないデータ量で近似的に記録さ

れたコンピュータ読み取り可能な記録媒体を提供することを第4の目的とする。

#### 【0014】

##### 【課題を解決するための手段】

本発明による生物学的物質の配列情報の記録方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さが $m$ ビット（ $m$ は16以上の整数）の部分データ（ $A(i, j)$ ）に分割し、複数列のその部分データに各行毎にその非配列方向にガロア体 $GF(2^m)$ 上の第1の演算を施して第1組のパリティ情報（ $B1(i)$ ， $B2(i)$ ， $B3(i)$ ）を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体 $GF(2^m)$ 上の第2の演算を施して第2組のパリティ情報（ $C1(j)$ ， $C2(j)$ ， $C3(j)$ ）を求め、その第1組及び第2組のパリティ情報でその生物学的物質の配列を表すものである。

#### 【0015】

斯かる本発明によれば、その生物学的物質（Biological materials）としては、例えば一列のヌクレオチド又は一列のアミノ酸が考えられる。前者の一列のヌクレオチドは、例えば或るDNA（deoxyribonucleic acid）を構成する1対の重合体の鎖の一方の鎖の少なくとも一部、或るRNA（ribonucleic acid）を構成する1列の重合体の鎖の少なくとも一部、又は或る遺伝子の少なくとも一部である。その一列のヌクレオチドの配列は、各ヌクレオチドに含まれる塩基の配列ともみなすことができる。一方、後者の一列のアミノ酸は、例えば或るタンパク質を構成するアミノ酸の配列の少なくとも一部である。

#### 【0016】

その生物学的物質の全体の個数を $NT$ として、各生物学的物質をそれぞれ1文字（例えばアルファベット）で表すものとする、その生物学的物質の配列に対応するテキストデータの全体の量は、例えばアスキーコード（ASCII code）（ANSI形式）では $NT$ バイトになり、ユニコード（Unicode）では $2 \cdot NT$ バイトになる。なお、配列を見易くするためのスペース、数字、及び改行などのコード

は無視している。そして、例えば図7の例において、テキストデータを配列方向に $N$ 個 ( $i = 1 \sim N$ ) で、非配列方向に $M$ 個 ( $j = 1 \sim M$ ) の部分テキストデータ  $T(i, j)$  に分割し、図8に示すように、各部分テキストデータ  $T(i, j)$  をそれぞれ $m$ ビットの部分データ  $A(i, j)$  に変換する。 $m$ ビットの部分データ  $A(i, j)$  は、それぞれ $n$ 個 (図8の例では $n = 16$ ) の連続する生物学的物質の配列を表している。

## 【0017】

この場合、最も簡単な方法としては、部分データ  $A(i, j)$  として部分テキストデータ  $T(i, j)$  そのものを数値データとみなしたデータを使用すればよい。即ち、テキストデータがアスキーコードで記録されている場合には、部分データ  $A(i, j)$  としてはそのアスキーコードを使用すればよい。また、テキストデータがユニコードで記録されている場合には、各文字をそれぞれの2バイトのコードの上位1バイトで表したものを部分データ  $A(i, j)$  としてもよい。但し、処理対象のデータ量を少なくするためには、各生物学的物質を表す文字を例えば6ビット以下の数値データに変換する変換テーブル (所定の規則) を用いて、部分テキストデータ  $T(i, j)$  を変換して得られる数値データを部分データ  $A(i, j)$  とすることが望ましい。

## 【0018】

次に、 $m$ ビットの各部分データ  $A(i, j)$  を非配列方向、及び配列方向に演算することによって、各行及び各列の配列情報を近似的に表すデータを算出する。このためには、 $m$ ビットのデータを加減乗除の対象にできる体 (Field) が必要であり、本発明ではそのためにガロア体 (拡大ガロア体)  $GF(2^m)$  を用いる。ガロア体  $GF(2^m)$  を用いた場合には、各行又は各列毎に $m$ ビットの部分データ  $A(i, j)$ 、及び必要に応じて $m$ ビットの係数を用いて所定の加減乗除演算 (第1又は第2の演算) を行ったときに得られる一つの情報 (これを「パリティ情報」と呼ぶ) が $m$ ビットであるため、配列情報を少ないデータ量で簡潔に記録できる利点がある。

## 【0019】

2を法とする数 (0及び1) で表される体を $Z_2$  とすると、ガロア体  $GF(2$

$m$  ) 上の演算は、体  $Z_2$  上の係数を持つ  $m$  次の既約多項式  $GF(X)$  を用いて定義することができる。即ち、2つの  $m$  ビットの部分データ  $A(i, j)$  ,  $A(i', j')$  をそれぞれ2進数表示で  $(a_{m-1} a_{m-2} \cdots a_1 a_0)$  ,  $(b_{m-1} b_{m-2} \cdots b_1 b_0)$  とすると ( $a_k, b_k$  は0又は1)、これらをそれぞれ次のように  $(m-1)$  次以下の多項式  $AF(X)$  ,  $BF(X)$  に変換する。

【0020】

$$AF(X) = a_{m-1} \cdot X^{m-1} + a_{m-2} \cdot X^{m-2} + \cdots + a_1 \cdot X + a_0 \cdots (1)$$

$$BF(X) = b_{m-1} \cdot X^{m-1} + b_{m-2} \cdot X^{m-2} + \cdots + b_1 \cdot X + b_0 \cdots (2)$$

この場合、ガロア体  $GF(2^m)$  上で  $AF(X)$  と  $BF(X)$  とを加算する場合には、 $X$  の各次数  $k$  ( $k=0 \sim (m-1)$ ) において、係数  $a_k$  と係数  $b_k$  とを体  $Z_2$  上で加算すればよい。体  $Z_2$  上では加算と減算とは同じ結果になる。この結果、得られた多項式の係数を2進数表示で表したもの(ベクトル表示)が、部分データ  $A(i, j)$  ,  $A(i', j')$  をガロア体  $GF(2^m)$  上で加算した結果になる。これは、ビット毎に排他的論理和演算を行うのと同じ結果である。

【0021】

次に、ガロア体  $GF(2^m)$  上で  $AF(X)$  に  $BF(X)$  を乗算する場合には、先ず通常の乗算を行って積を求めた後、この積を既約多項式  $GF(X)$  で除算した後の余りの多項式  $CF(X)$  を次のように求める ( $c_k$  は0又は1)。これを既約多項式  $GF(X)$  を法(modulus)とする乗算と呼ぶ。この際にも  $X$  の各次数での係数の加算(減算)は体  $Z_2$  上で行われる。

【0022】

$$CF(X) = c_{m-1} \cdot X^{m-1} + c_{m-2} \cdot X^{m-2} + \cdots + c_1 \cdot X + c_0 \cdots (3)$$

この多項式  $CF(X)$  の係数を2進数表示で表したもの ( $c_{m-1} c_{m-2} \cdots c_1 c_0$ ) が、部分データ  $A(i, j)$  ,  $A(i', j')$  をガロア体  $GF(2^m)$  上で乗算した結果になる。また、任意の  $m$  ビットの係数を  $\beta$  とすると、係数  $\beta$  も(2)式と同様の  $(m-1)$  次以下の多項式  $DF(X)$  で表される。従って、例えば部分データ  $A(i, j)$  に係数  $\beta$  を乗算する場合には、(1)式の多項式  $AF(X)$  と多項式  $DF(X)$  との積を既約多項式  $GF(X)$  を法として計算す

ればよい。また、例えば部分データ  $A(i, j)$  を係数  $\beta$  で除算する場合には、部分データ  $A(i, j)$  に  $\beta$  の逆元  $\beta^{-1}$  を乗算すればよい。

【 0 0 2 3 】

従って、 $m$  ビットの全てのデータ（全ての部分データ  $A(i, j)$  が含まれる）は、ガロア体  $GF(2^m)$  上のベクトル表示での元とみなすことができ、 $m$  ビットのデータは、多項式表示では、(1) 式のような  $(m-1)$  次以下の多項式で表すことができる。また、生物学的物質の配列（文字列）を部分データ  $A(i, j)$  に対応させる変換テーブル（所定の規則）の逆変換を用いて、必要に応じてそのベクトル表示の  $m$  ビットのデータを文字列に変換することによって、そのデータに対応する生物学的物質の配列が得られる。

【 0 0 2 4 】

そして、本発明では、例えば図 8 に示すように、部分データ  $A(i, j)$  が配列方向に  $N$  個 ( $i = 1 \sim N$ ) で、非配列方向に  $M$  個 ( $j = 1 \sim M$ ) で配列され、各行毎に第 1 組のパリティ情報 ( $B1(i), B2(i), B3(i)$ ) が得られ、各列毎に第 2 組のパリティ情報 ( $C1(j), C2(j), C3(j)$ ) が得られる。これら 2 組のパリティ情報の内の 1 つのパリティ情報（例えば  $B1(1)$ ）はそれぞれ 1 つの部分データ  $A(i, j)$  と同じ  $m$  ビットのデータで表される。

【 0 0 2 5 】

この場合の部分データ  $A(i, j)$  の全体のデータ量  $DT1$  は、以下のようになる。

$$DT1 = m \cdot N \cdot M \text{ (ビット)} \quad \dots (4)$$

また、第 1 組及び第 2 組のパリティ情報が、それぞれ  $e$  個 ( $e$  は 1 以上の整数) のパリティ情報を含むとすると、パリティ情報全体のデータ量  $DS1$  は、以下のようになる。なお、 $e$  個のパリティ情報を含む場合には、各行及び各列において、それぞれ  $e$  個までの部分データ  $A(i, j)$  を復元できる。

【 0 0 2 6 】

$$DS1 = m \cdot e \cdot (N + M) \text{ (ビット)} \quad \dots (5)$$

従って、例えば生物学的物質が DNA を構成するヌクレオチドであるとして、

仮に  $N = 64$ ,  $M = 128$ ,  $e = 2$  とすると、(4) 式及び (5) 式よりデータ量  $DT1$ ,  $DS1$  は以下ようになる。

$$DT1 = m \cdot 8192 \text{ (ビット)} \quad \dots (6)$$

$$DS1 = m \cdot 384 \text{ (ビット)} \doteq DT1 / 20 \quad \dots (7)$$

従って、パリティ情報のデータ量は、部分データ  $A(i, j)$  全体のデータ量のほぼ  $1/20$  程度に少なくできる。この場合、例えば人間の 1 本の染色体の DNA の配列は、50M バイト ~ 250M バイト程度のテキストデータで表されるため、予めそのテキストデータを 500 個 ~ 2500 個程度のブロックに分割し、各ブロック毎に 2 組のパリティ情報を求めることによって、全部のパリティ情報のデータ量はそのテキストデータのほぼ  $1/20$  程度、即ち 2.5M バイト ~ 12.5M バイト程度に少なくできる。また、その部分データ  $A(i, j)$  が、例えばテキストデータを  $1/f$  に小さくしたデータ量である場合には、パリティ情報も更に  $1/f$  だけ少なくすることができる。

#### 【0027】

本発明によれば、元の生物学的物質の配列情報を近似的に表す情報（パリティ情報）を、元のテキストデータよりも少ないデータ量のファイルに記録することができる。従って、記録媒体として、DVD-ROM のような大容量の媒体の他に、CD-ROM、及びフラッシュROM のような小容量でも通常のコンピュータで手軽に再生できる媒体を使用できる。更に、少ないデータ量の配列情報であれば、通信回線を介して短時間に送信できるため、そのパリティ情報は、例えば携帯電話システムなどを介してユーザに安価に供給することも可能となる。

#### 【0028】

そして、第 1 組のパリティ情報、及び第 2 組のパリティ情報を用いることによって、ユーザ側では、2 つの生物学的物質の配列の相違する部分を容易に特定することができると共に、相違する部分の個数が各行、又は各列で  $e$  個以下である場合には、パリティ情報を用いて相違する部分の配列の復元を行うことも可能となる。

#### 【0029】

なお、テキストデータが記録されたファイルが通常の圧縮技術（ZIP ファイ



ル、LHAファイル等)で圧縮できるように、本発明のパリティ情報が記録されたファイルも更に通常の圧縮技術を用いて圧縮して記録できることは言うまでもない。但し、圧縮されたファイルを使用する場合には、解凍作業が必要になり、最終的には元のファイルを復元する必要があるため、元のファイル自体のデータ量を減らしておくことは極めて有効である。

【0030】

次に、上記の本発明において、そのガロア体 $GF(2^m)$ 上の生成元を $\alpha$ としたとき、一例として、その第1組のパリティ情報は、複数行の各行のその部分データ $(A(i, j))$ にそれぞれその非配列方向に順次 $\alpha^{sp}$ ,  $\alpha^{s(p+1)}$ ,  $\alpha^{s(p+2)}$ , ...,  $\alpha^{s(p+dp)}$  ( $s$ は0以上の整数、 $p$ は0以上の整数、 $dp$ は1以上の整数)を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第2組のパリティ情報は、複数列の各列のその部分データ $(A(i, j))$ にそれぞれその配列方向に順次 $\alpha^{tq}$ ,  $\alpha^{t(q+1)}$ ,  $\alpha^{t(q+2)}$ , ...,  $\alpha^{t(q+dq)}$  ( $t$ は0以上の整数、 $q$ は0以上の整数、 $dq$ は1以上の整数)を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。

【0031】

この場合、 $p = q = 0$ とすると、第1組のパリティ情報 $B1(i)$ 、及び第2組のパリティ情報 $C1(j)$ は、それぞれガロア体 $GF(2^m)$ 上の次の演算によって計算される。(8)式の $\Sigma$ は $j$ について1~ $M$ までの和を表し、(9)式の $\Sigma$ は $i$ について1~ $N$ までの和を表している。

$$B1(i) = \Sigma \alpha^{s(j-1)} \cdot A(i, j) = A(i, 1) + \alpha^s \cdot A(i, 2) + \alpha^{2s} \cdot A(i, 3) + \dots + \alpha^{(M-1)s} \cdot A(i, M) \quad \dots (8)$$

$$C1(j) = \Sigma \alpha^{t(i-1)} \cdot A(i, j) = A(1, j) + \alpha^t \cdot A(2, j) + \alpha^{2t} \cdot A(3, j) + \dots + \alpha^{(N-1)t} \cdot A(N, j) \quad \dots (9)$$

そして、(8)式、(9)式で $s = t = 0$ とすると、パリティ情報 $B1(i)$ 、 $C1(j)$ は、それぞれ部分データ $A(i, j)$ のガロア体 $GF(2^m)$ 上の和、即ち各行又は各列で部分データ $A(i, j)$ に排他的論理和演算を施して得られる値を示す。従って、簡単な演算で、各行及び各列の配列の近似的な情報を

求めることができる。但し、この場合には、各行又は各列で2つの部分データ  $A(i, j)$  が入れ替わったような配列に対しても、パリティ情報  $B1(i)$ ,  $C1(j)$  は同じ値になってしまう。

## 【0032】

次に、 $s = t = 1$  とすると、パリティ情報  $B1(i)$ ,  $C1(j)$  は、それぞれ各行又は各列で部分データ  $A(i, j)$  に  $1, \alpha, \alpha^2, \alpha^3, \dots$  を乗算して得られる積の和を示す。この場合、各行又は各列で2つの部分データ  $A(i, j)$  が入れ替わったような配列に対しても、パリティ情報  $B1(i)$ ,  $C1(j)$  は異なった値となるため、例えば2つの生物学的物質の配列間の相違する部分をより正確に特定することができる。そして、或る  $s (\neq 0)$  (又は  $t (\neq 0)$ ) の値において、そのように各行又は各列で2つの部分データに常に異なる係数を乗ずるためには、係数  $\alpha^{s(j-1)}$  (又は  $\alpha^{t(i-1)}$ ) は互いに異なる必要がある。そのためには、 $\alpha$  をガロア体  $GF(2^m)$  上の生成元として、各行及び各列の部分データ  $A(i, j)$  の個数を  $(2^m - 1) / s$  (又は  $(2^m - 1) / t$ ) 以下とすればよい。即ち、 $\alpha$  を生成元とすることによって、扱う生物学的物質の配列を最も大きくすることができる。

## 【0033】

また、各行及び各列において、それぞれ1つのパリティ情報を用いることによって、2つの生物学的物質の配列を比較する場合に、各行及び各列における1つの部分データ  $A(i, j)$  の相違部を正確に復元することができる。従って、例えば遺伝子中の一つの塩基（ヌクレオチド）だけが異なる SNP（一塩基変位多型：Single Nucleotide Polymorphism）は本発明によって容易に検出できると共に、それに対応する正常な配列も容易に復元できる。

## 【0034】

更に、各行及び各列における複数個  $s'$  及び  $t'$  の部分データ  $A(i, j)$  の相違部を正確に復元するためには、その第1組のパリティ情報 ( $B1(i)$ ,  $B2(i)$ ,  $B3(i)$ ) は、その複数行の各行毎にその整数  $s$  について互いに異なる複数 ( $s'$  個) の値で求めた複数の和を含み、その第2組のパリティ情報 ( $C1(j)$ ,  $C2(j)$ ,  $C3(j)$ ) は、その複数列の各列毎にその整数  $t$  に

ついて互いに異なる複数 ( $t'$  個) の値で求めた複数の和を含むようにすればよい。その相違部を復元するためには、ガロア体  $GF(2^m)$  上で  $s'$  元 ( $t'$  元) の 1 次連立方程式を解けばよい。

#### 【 0 0 3 5 】

また、本発明において、そのガロア体  $GF(2^m)$  を規定する整数  $m$  は 64 の倍数であることが望ましい。最近のコンピュータにはデータの処理単位が 64 ビットであるタイプが増加しているため、整数  $m$  を 64 の倍数とすることによって、効率的にパリティ情報を算出することができる。

また、本発明において、その生物学的物質の配列を基準配列として、この基準配列のその 2 組のパリティ情報に対応させて、検査対象の生物学的物質の配列についてその 2 組のパリティ情報を求め、その 4 組のパリティ情報よりその基準配列に対するその検査対象の生物学的物質の配列の相違部を求めるようにしてもよい。このように基準配列の 2 組のパリティ情報と、検査対象の 2 組のパリティ情報とを比較するのみで、相違部の位置を容易に検出できると共に、相違部が各行及び各列で所定個数以下であれば、4 組のパリティ情報と検査対象の生物学的物質の配列とから、連立方程式を解くことによって、基準配列の相違部のデータを正確に復元することもできる。

#### 【 0 0 3 6 】

次に、本発明の生物学的物質の配列情報の記録装置は、その生物学的物質の配列情報を読み取る配列読み取り装置 (4) と、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数列で、かつその配列方向に交差する非配列方向に複数列の長さが  $m$  ビット ( $m$  は 16 以上の整数) の部分データ ( $A(i, j)$ ) に分割するデータ配列手段 (10, ステップ 105) と、複数列のその部分データに各行毎にその非配列方向にガロア体  $GF(2^m)$  上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体  $GF(2^m)$  上の第 2 の演算を施して第 2 組のパリティ情報を求める演算手段 (10, ステップ 106) と、その第 1 組及び第 2 組のパリティ情報を記録媒体に記録する記録手段 (15) とを有する

ものである。これによって、本発明のヌクレオチドやアミノ酸などの生物学的物質の配列情報の記録方法が実施できる。

#### 【 0 0 3 7 】

この本発明の記録装置において、そのガロア体  $GF(2^m)$  上の生成元を  $\alpha$  としたとき、一例としてその第 1 組のパリティ情報は、複数行の各行のその部分データ  $(A(i, j))$  にそれぞれその非配列方向に順次  $\alpha^{sp}$ ,  $\alpha^{s(p+1)}$ ,  $\alpha^{s(p+2)}$ , ...,  $\alpha^{s(p+dp)}$  ( $s$  は 0 以上の整数、 $p$  は 0 以上の整数、 $dp$  は 1 以上の整数) を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第 2 組のパリティ情報は、複数列の各列のその部分データにそれぞれその配列方向に順次  $\alpha^{tq}$ ,  $\alpha^{t(q+1)}$ ,  $\alpha^{t(q+2)}$ , ...,  $\alpha^{t(q+dq)}$  ( $t$  は 0 以上の整数、 $q$  は 0 以上の整数、 $dq$  は 1 以上の整数) を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。この場合、部分データ  $(A(i, j))$  に乗ずる係数は、生成元  $\alpha$  だけから計算できるため、演算が単純化される。

#### 【 0 0 3 8 】

また、本発明の記録媒体は、生物学的物質の配列情報を記録したコンピュータ読み取り可能な記録媒体 (16) であって、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さが  $m$  ビット ( $m$  は 16 以上の整数) の部分データに分割し、複数行のその部分データに各行毎にその非配列方向にガロア体  $GF(2^m)$  上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体  $GF(2^m)$  上の第 2 の演算を施して第 2 組のパリティ情報を求め、その生物学的物質の配列に関する情報を、その第 1 組及び第 2 組のパリティ情報として記録したものである。

#### 【 0 0 3 9 】

本発明によれば、ヌクレオチドやアミノ酸などの生物学的物質の配列情報を近似的に表すパリティ情報を少ないデータ量でその記録媒体に記録できるため、記録媒体として CD-ROM, CD-R, フラッシュ ROM などの記録容量は比較

的が少ないが、使い勝手の良い媒体をも使用できる。

この場合、その生物学的物質の配列に対応するそのテキストデータ、又はこのテキストデータに対応するその数値データの40ビット以上の長さの数学的な要約値 (message digest) を更にその記録媒体に記録することが望ましい。

#### 【0040】

その数学的な要約値は、その生物学的物質の配列に対応するテキストデータ又は数値データに例えばMD5ハッシュ関数 (要約値は128ビット)、又はSHS (Secure Hash Standard) ハッシュ関数 (要約値は160ビット) などのハッシュ関数の演算を施して得られるものである。その要約値を用いることによって、比較対象の2つの生物学的物質の膨大な配列が一致するかどうかを高い確率で極めて容易に確認することができる。また、パリティ情報を用いて相違する部分データの復元を行った後に、要約値を比較することによって、データが完全に復元できたかを確認することができる。その要約値が40ビット以上であれば、例えば全人類のDNAのヌクレオチドの配列情報をほぼ互いに衝突することなく表すことができる。

#### 【0041】

この場合、ガロア体  $GF(2^m)$  を決定する整数  $m$  が64の倍数であるときには、要約値が64ビットの倍数となるハッシュ関数 (例えばMD5ハッシュ関数) を用いることが望ましい。演算を効率的に実行できるからである。

次に、本発明の生物学的物質の配列情報の供給方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを保持する供給者 (2A) が、そのテキストデータ、又はこれに対応するその数値データを第1ファイル (19) に記録して保持する第1ステップ (ステップ104) と、その第1ファイルに記録されているそのテキストデータ、又はこのテキストデータに対応するその数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さが  $m$  ビット ( $m$  は16以上の整数) の部分データに分割し、複数行のその部分データに各行毎にその非配列方向にガロア体  $GF(2^m)$  上の第1の演算を施して第1組のパリティ情報 ( $B1(i) \sim B3(i)$ ) を求めると共に、複

数列のその部分データに各列毎にその配列方向にガロア体  $GF(2^m)$  上の第2の演算を施して第2組のパリティ情報 ( $C1(j) \sim C3(j)$ ) を求める第2ステップ (ステップ105, 106) と、その供給者が、その第1組及び第2組のパリティ情報を第2ファイル (20) に記録して保持する第3ステップ (ステップ107) と、その生物学的物質の配列情報のユーザ (2B) が、通信回線 (1) を介してその供給者よりその第2ファイルに記録されているその2組のパリティ情報を受け取る第4ステップ (ステップ110, 129) とを有するものである。

## 【0042】

この供給方法は、上記の本発明の生物学的物質の配列情報の記録方法を、その配列情報を供給 (販売) する際のビジネスモデルに適用したものである。即ち、本発明のビジネスモデルでは、或る生物XのDNAのヌクレオチド、又はタンパク質のアミノ酸などの生物学的物質の配列を最初に解読した供給者は、その配列のテキストデータ (又はこれを変換した数値データ) より、その配列情報を少ないデータ量で近似するパリティ情報を算出し、これをその通信回線を介してユーザに供給する。上述のように、一例としてパリティ情報は、元のテキストデータの  $1/20$  程度のデータ量であるため、そのパリティ情報はその通信回線を介して短時間で受信することができる。

## 【0043】

本発明の供給方法においては、更にそのユーザが、その2組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内のその供給者の生物学的物質の配列情報との相違部を特定する第5ステップ (ステップ130, 131) と、この相違部の配列の復元ができない場合に、そのユーザがその通信回線を介してその供給者よりその第1ファイルに記録されているそのテキストデータ、又はその数値データの内のその配列の復元ができない部分の配列情報を受け取る第6ステップ (ステップ135) とを有することが望ましい。

## 【0044】

このようにユーザ側で、パリティ情報だけで検査対象の配列内の供給者の配列との相違部の特定、及び復元ができる場合には、それ以上の配列情報を購入する

必要が無い。一方、相違部が多く存在し、パリティ情報のみでは全部の正確なデータが復元できない場合には、例えば復元できない部分のテキストデータ（又は数値データ）のみを購入することによって、通信回線を介して必要な配列情報を短時間に購入できる。従って、通信回線として、携帯電話システムのような比較的低速の通信回線も使用できる。

## 【0045】

また、本発明の供給方法においては、そのガロア体  $GF(2^m)$  上の生成元を  $\alpha$  としたとき、一例として、その第1組のパリティ情報は、複数行の各行のその部分データにそれぞれその非配列方向に順次  $\alpha^{sp}$ ,  $\alpha^{s(p+1)}$ ,  $\alpha^{s(p+2)}$ , ...,  $\alpha^{s(p+dp)}$  ( $s$  は0以上の整数、 $p$  は0以上の整数、 $dp$  は1以上の整数) を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第2組のパリティ情報は、複数列の各列のその部分データにそれぞれその配列方向に順次  $\alpha^{tq}$ ,  $\alpha^{t(q+1)}$ ,  $\alpha^{t(q+2)}$ , ...,  $\alpha^{t(q+dq)}$  ( $t$  は0以上の整数、 $q$  は0以上の整数、 $dq$  は1以上の整数) を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。

## 【0046】

これらのパリティ情報を用いることによって、そのユーザは、SNP（一塩基変位多型）などを容易に検出することができる。

また、その供給方法においては、その供給者は、その生物学的物質の配列の長さの情報、及びその配列を表すテキストデータ又はその数値データの数学的な要約値の情報をその通信回線を介して閲覧可能な状態にしておき、そのユーザは、その第4ステップの前にその通信回線を介してその配列の長さの情報及びその数学的な要約値の情報を閲覧する（ステップ121）ことが望ましい。

## 【0047】

この場合、その供給者は、その生物Xの生物学的物質の配列のテキストデータ（又はこれを変換した数値データ）よりハッシュ関数によって算出した要約値（message digest）を例えばインターネット上で閲覧可能にする。これによって、その供給者は、そのテキストデータ自体を公開することなく、最初にその生物Xの生物学的物質の配列を解読したことを主張できる。更に、ユーザが同じ配列情

報を異なる供給者から誤って購入することも防止できる。

【 0 0 4 8 】

また、或るユーザが、その供給者よりその生物学的物質の配列情報を購入した後、購入した配列情報よりそのハッシュ関数によって要約値を算出し、その配列の長さも求める。そして、この配列の長さ、及び要約値をインターネット上で公開されている値と比較することによって、購入した配列情報が正確なものであるかどうかを極めて高い確率で確認できる。

【 0 0 4 9 】

この場合、一例として、その数学的な要約値は、40ビット以上で192ビット以下のデータであり、その供給者は、更にその生物学的物質の所定の一部の配列の情報をその通信回線を介して閲覧可能な状態にしておくことが望ましい。その要約値、及びその配列の長さの他に、そのように例えばその配列の先頭の8個程度、及び後端の8個程度の配列を比較することによって、同一性の確認をより高精度に行うことができる。

【 0 0 5 0 】

【発明の実施の形態】

以下、本発明の実施の形態の一例につき図面を参照して説明する。本例は、所定のDNA（デオキシリボ核酸：deoxyribonucleic acid）のヌクレオチド（生物学的物質）の配列情報をコンピュータシステムで処理する場合に、本発明を適用したものである。

【 0 0 5 1 】

図1は、本例のコンピュータシステム2Aの概略構成を示し、この図1において、コンピュータシステム2Aの中心は、CPU（中央演算処理ユニット）、RAM、ROM等のメモリ、及びハードディスク装置等の記憶装置等からなる情報処理装置10である。情報処理装置10には、ビデオRAM（VRAM）11を介してCRTディスプレイよりなる表示装置12が接続されると共に、I/Oユニット（入出力装置）14を介して、記録可能なCD-Recordableディスク（以下、「CD-R」と言う）16に対するデータの書き込み、及びCD-RやCD-ROMからのデータの読み込みを行うことができるCD-R/RWドライブ1



5 が接続されている。情報処理装置 1 0 には、I / O ユニット 1 4 を介して更に大容量の記憶装置としての記憶容量が数 1 0 0 G バイト程度の磁気ディスク装置 1 7 が接続されている。

【 0 0 5 2 】

本例の情報処理装置 1 0 中のハードディスク装置には、予め C D - R / R W ドライブ 1 5 を介してオペレーティングシステム、及び後述のように D N A の配列情報を処理するためのアプリケーション・プログラムがインストールされている。また、C D - R 1 6 が本発明の記録媒体に対応しているが、記録媒体としては、C D - R や C D - R O M の他に、フラッシュ R O M 、フレキシブルディスク、光磁気ディスク ( M O ) 、デジタルビデオディスク ( D V D ) 、又はハードディスク装置 ( 例えばインターネットを介して接続できるサーバに備えられたもの ) 等を使用することができる。

【 0 0 5 3 】

情報処理装置 1 0 には更に、文字情報の入力装置としてのキーボード 1 3 、ポインティング・デバイス ( 入力装置 ) としての光学式のマウス 2 0 4 、及びルータ ( 又はモデム等でもよい ) よりなる通信制御ユニット 1 8 が接続されている。マウス 2 0 4 は、表示装置 1 2 の表示画面上のカーソルの位置を指定する信号を発生する変位信号発生部 2 0 7 、選択すべき情報を指定する信号や各種コマンド等を発生するための左スイッチ 2 0 4 a 及び右スイッチ 2 0 4 b ( 信号発生装置 ) を備えている。情報処理装置 1 0 、 V R A M 1 1 、表示装置 1 2 、キーボード 1 3 、マウス 2 0 4 、 I / O ユニット 1 4 、 C D - R / R W ドライブ 1 5 、磁気ディスク装置 1 7 、及び通信制御ユニット 1 8 等よりコンピュータシステム 2 A が構成されている。オペレーティングシステムとして本例では W i n d o w s ( Microsoft Corporation の登録商標 ) を使用している。なお、オペレーティングシステムとして、それ以外の U N I X ( X / O p e n の登録商標 ) 、 O S / 2 ( I B M Corporation の登録商標 ) 、 M a c O S ( Apple Computer の登録商標 ) 、又は L i n u x ( Linus Torvalds の商標又は登録商標 ) 等を使用する場合にも本発明が適用できることは言うまでもない。

【 0 0 5 4 】

そして、コンピュータシステム 2 A（情報処理装置 1 0）は、通信制御ユニット 1 8 を介して一般電話回線よりなる通信ネットワーク 1 に接続され、通信ネットワーク 1 には各種コンテンツのプロバイダ 3、及び別のコンピュータシステム 2 B、及び不図示の多くのサーバやコンピュータシステムが接続されている。また、本例のコンピュータシステム 2 A、2 B 及びプロバイダ 3 は、通信ネットワーク 1 を介するインターネットによって相互に接続されている。この場合、コンピュータシステム 2 A の所有者が DNA 情報の供給者（販売者）であり、コンピュータシステム 2 B の所有者がその DNA 情報のユーザ（購入者）である。そして、後者のコンピュータシステム 2 B には、予め前者のコンピュータシステム 2 A と同様の DNA の配列情報を処理するためのアプリケーション・プログラムがインストールされている。

#### 【 0 0 5 5 】

さて、本例のコンピュータシステム 2 A の情報処理装置 1 0 には、I / O ユニット 1 4 を介して、生物学的物質としての DNA 中の一列のヌクレオチドの配列（又は塩基の配列）を読み取るための配列読み取り装置としてのシーケンサー（DNA Sequencer）4 が接続されている。シーケンサー 4 は、一例としてサンガーの方法（Sanger method）によって DNA を構成する 1 対の重合体の鎖の一方の鎖のヌクレオチドの配列を読み取る。サンガーの方法は、例えば文献 1（Maxim D. Frank-Kamenetskii: Unraveling DNA (the most important molecule of life, revised and updated), translated by Lev Liapin, Chapter 6 (pp.59-70) (Perseus Books, 1997)）に開示されている。シーケンサー 4 は、読み取った一列のヌクレオチドの配列をテキストデータ形式で内部の大容量の記憶装置に記憶すると共に、情報処理装置 1 0 からの要求に応じて、その記憶装置中の所定のヌクレオチドの配列のテキストデータを I / O ユニット 1 4 を介して情報処理装置 1 0 に供給する。これに対して情報処理装置 1 0 は、DNA の配列情報を処理するためのアプリケーション・プログラムに基づいて以下の処理を行う。なお、シーケンサー 4 の代わりに、DNA 及び RNA（リボ核酸：ribonucleic acid）等の核酸を構成する一列のヌクレオチドの配列（又は塩基の配列）の情報のデータベースを接続してもよい。

## 【 0 0 5 6 】

先ず、本例の情報処理装置 1 0 の第 1 の基本的な処理動作につき説明する。情報処理装置 1 0 は、シーケンサ 4 から供給される所定の DNA のヌクレオチドの配列を示すテキストデータ（本例ではアスキーコード（ANSI 形式）を用いる）を磁気ディスク装置 1 7 中のマスターファイル 1 9 にそのまま記録すると共に、そのテキストデータをよりデータ量の少ない数値データに変換し、この変換後の数値データを磁気ディスク装置 1 7 中のワーキングファイル 2 0 に記録する。なお、以下の説明において、2 進数表示の数  $k$  は  $\text{bin}(k)$  で、1 6 進数表示の数  $k$  は  $\text{hex}(k)$  で表すものとする。

## 【 0 0 5 7 】

この場合、DNA は 4 種類のヌクレオチドより構成されており、シーケンサ 4 から供給されるテキストデータ中では、塩基としてアデニン（adenine）を含むヌクレオチド、グアニン（guanine）を含むヌクレオチド、シトシン（cytosine）を含むヌクレオチド、及びチミン（thymine）を含むヌクレオチドがそれぞれ文字 A, G, C, 及び T で表されている。そして、文字 A, G, C, 及び T には、データ上ではそれぞれ  $\text{hex}(41)$ ,  $\text{hex}(47)$ ,  $\text{hex}(43)$ ,  $\text{hex}(54)$  よりなる 1 バイト（8 ビット）のアスキーコードが割り当てられている。また、RNA の場合には、チミンを含むヌクレオチドの代わりにウラシル（uracil）を含むヌクレオチドが、文字 U（ $\text{hex}(55)$ ）で表されている。従って、 $n$  個のヌクレオチドの配列を示すテキストデータのデータ量は  $n$  バイトとなる。なお、それらの  $n$  個のヌクレオチドの配列は、 $n$  個の塩基（アデニン、グアニン、シトシン、チミン（又はウラシル））の配列ともみなすことができる。

## 【 0 0 5 8 】

本例ではそのテキストデータを、情報量を少なくすることなく最も少ないデータ量で表すために、DNA 中の 4 種類のヌクレオチドを互いに異なる 2 ビットのデータで表す。この際に、DNA においては、1 対の塩基（アデニン及びチミン）が互いに相補的であり、別の 1 対の塩基（グアニン及びシトシン）が互いに相補的である。そこで、相補的な塩基を含む 1 対のヌクレオチドを互いに相補的であるとして、1 対の互いに相補的なヌクレオチド、即ちアデニンを含むヌクレオ

チド及びチミンを含むヌクレオチドに、互いにビット反転の関係にある1対のデータを割り当て、別の1対の互いに相補的なヌクレオチド、即ちグアニンを含むヌクレオチド及びシトシンを含むヌクレオチドに、互いにビット反転の関係にある別の1対のデータを割り当てる。本例ではそのデータの割り当てとして表1（変換テーブル）を用いる。なお、表1は、ヌクレオチドの配列を示すテキストデータ中の文字A, T（又はU）, G, C, をそれぞれbin(00), bin(11), bin(01), bin(10) で置換することを意味している。

【0059】

《表1》

ヌクレオチド	2ビットのデータ
アデニンを含むヌクレオチド（A）	bin(00)
チミン（ウラシル）を含むヌクレオチド（T又はU）	bin(11)
グアニンを含むヌクレオチド（G）	bin(01)
シトシンを含むヌクレオチド（C）	bin(10)。

【0060】

なお、本例では各ヌクレオチドを2ビットのデータで表しているが、これは各塩基を2ビットのデータで表すのと等価である。また、データの割り当ては表1には限定されず、例えばチミンを含むヌクレオチドをbin(00)、アデニンを含むヌクレオチドをbin(11) とするか、又はグアニンを含むヌクレオチドをbin(10)、シトシンを含むヌクレオチドをbin(01) としてもよい。それ以外に、アデニンを含むヌクレオチド及びチミンを含むヌクレオチドに、1対のデータbin(01), bin(10) を割り当て、グアニンを含むヌクレオチド及びシトシンを含むヌクレオチドに1対のデータbin(00), bin(11) を割り当てるようにしてもよい。また、RNAの場合には、チミンを含むヌクレオチドに割り当てられているデータをウラシルを含むヌクレオチドに割り当てて、それ以外のヌクレオチドにはDNAのヌクレオチドと同じデータを割り当てればよい。

【0061】

本例では図2に示すDNA分子5のヌクレオチドの配列情報を扱うものとする。その配列情報は、NCBI（The National Center for Biotechnology Inform

ation) より提供されているウェブサイト 1 (<ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/>) より入手した大腸菌 (*Escherichia coli*: *E. coli*) の DNA の一列のヌクレオチドの配列の一部である。

【 0 0 6 2 】

図 2 において、DNA 分子 5 は、1 対の重合体の鎖 6 A、6 B (二重らせん) より構成され、一方の重合体の鎖 6 A は、アデニンを含むヌクレオチド 7 A、グアニンを含むヌクレオチド 7 G、シトシンを含むヌクレオチド 7 C、及びチミンを含むヌクレオチド 7 T よりなる 4 種類のヌクレオチドの配列であり、他方の重合体の鎖 6 B は、鎖 6 A に対して相補的なヌクレオチドの配列である。この際に、図 1 の情報処理装置 1 0 には一方の重合体の鎖 6 A の配列を示すテキストデータ、即ち "AGCTTT..." の文字列のデータが供給される。それに対して、情報処理装置 1 0 は、そのテキストデータ中の文字 A、G、C、T を表 1 の変換テーブルに基づいて順次 2 ビットのデータに変換することによって、数値データとしてのバイナリーデータ BNA (=bin(0001101111...)) を得る。そして、このバイナリーデータ BNA が図 1 の磁気ディスク装置 1 7 のワーキングファイル 2 0 に記録される。そのバイナリーデータ BNA のデータ量は、元のテキストデータの 1/4 となっている。

【 0 0 6 3 】

この場合、そのワーキングファイル 2 0 の先頭の所定数のバイトの領域に、例えばその配列が DNA 又は RNA のどちらかを示すデータ (即ち、bin(11) を文字 T 又は文字 U の何れに解釈するかを示すデータ)、ヌクレオチドの個数を示すデータ、及びその他の必要なデータを記録しておけばよい。また、そのワーキングファイル 2 0 の長さが 1 バイト (8 ビット) 単位で規定されている場合に、バイナリーデータ BNA の末尾で 1 バイトの端数のデータが生じたときには、予め定めておいたダミーデータを付加すればよい。これでもデータ量は殆ど増加しない。そして、一例としてユーザ (コンピュータシステム 2 B の所有者) から供給者 (コンピュータシステム 2 A の所有者) に対して図 2 の DNA 分子 5 の配列情報の購入希望が届いたときに、ワーキングファイル 2 0 のデータが通信ネットワーク 1 及び不図示のプロバイダを介して、電子メールの添付ファイルとしてコン

コンピュータシステム 2 B 側に供給される。この際に、そのワーキングファイル 2 0 のデータを更に圧縮ファイル（Z I P ファイル、又は L H A ファイル等）として送信してもよい。この際に、ワーキングファイル 2 0 のデータ量はもとのテキストデータのほぼ 1 / 4 であるため、元のテキストデータ（更に圧縮ファイルとした場合も同様）自体を送信する場合に比べて送信時間はほぼ 1 / 4 となり、供給者側及びユーザ側双方の通信コストが低減できる。

## 【 0 0 6 4 】

そして、ユーザ側で、その受信したワーキングファイル 2 0 のデータから図 2 の一方の重合体の鎖 6 A の配列のテキストデータを復元する場合には、コンピュータシステム 2 B において、ワーキングファイル 2 0 中のバイナリーデータ B N A を、表 1 を用いて文字 A, G, C, T（又は U）の何れかに順次逆変換すればよい。また、その際に例えば図 2 の他方の相補的な重合体の鎖 6 B のヌクレオチドの配列を示すテキストデータが必要になった場合には、コンピュータシステム 2 B において、図 2 に示すように、バイナリーデータ B N A のビット毎の反転操作を行って反転バイナリーデータ NOT(BNA) (=bin(1110010000...)) を得る。この反転バイナリーデータ NOT(BNA) は、他方の重合体の鎖 6 B のヌクレオチドの配列を示すテキストデータ（文字列” T C G A A A . . . ”）を表 1 に従って変換したバイナリーデータ B N B と同一である。従って、その反転バイナリーデータ NOT(BNA) を、表 1 を用いて文字 A, G, C, T（又は U）の何れかに順次逆変換するのみで、極めて高速に相補的な重合体の鎖 6 B の配列のテキストデータを得ることができる。この際に、通常のコンピュータにおいては、ビット毎の反転操作は、極めて高速に実行することができる。なお、そのビット毎の反転操作は、例えば bin(111111...) との排他的論理和演算で代用してもよい。

## 【 0 0 6 5 】

なお、ワーキングファイル 2 0 のデータを通信ネットワーク 1 を介してユーザ側に送信する代わりに、ワーキングファイル 2 0 の内容を C D - R / R W ドライブ 1 5 によって C D - R 1 6 に記録し、この C D - R 1 6 を郵送等によってユーザ側に供給してもよい。例えば一人の人間の全部の D N A の配列情報（ゲノム）は、テキストデータでは 3 G バイト程度になるが、これを表 1 を用いて本例の数

値データとしてのバイナリーデータに変換すると、3/4 Gバイト程度、即ち750 Mバイト程度になる。現在のCD-R、CD-ROMの容量は約650 Mバイトであるため、その750 Mバイト程度のバイナリーデータは例えば一部又は全部を圧縮ファイルとすることによって、余裕を持ってCD-R 16に記録することができる。これに対して、その750 Mバイト程度のデータを通信ネットワーク1を介して送信しようとする、現状でも送信時間がかかり過ぎる場合がある。

## 【0066】

また、一つのアミノ酸の種類は一系列の3個のヌクレオチドの配列、即ち一つの遺伝子コドン (codon) によって決定される。そこで、1つのアミノ酸に対応する3個のヌクレオチドをそれぞれ2ビットのデータで表したときに得られる6ビットのデータの中で、最も小さいデータでそのアミノ酸を表すようにしてもよい。この際に、個々のデータは、1バイト単位が扱い易いため、その6ビットのデータの前後に2ビットの識別データを付加して得られる1バイトのデータで1つのアミノ酸を表すようにしてもよい。これによって、ヌクレオチドとアミノ酸とで共通のコードを使用できる利点がある。

## 【0067】

次に、本例の情報処理装置10の第2の基本的な処理動作につき説明する。本例では、ヌクレオチドの配列を示す膨大な量のテキストデータ（又はこれを表1に基づいて変換して得られる数値データ）より、所定のハッシュ関数を用いて数学的な要約値 (message digest) を算出する。本例ではそのハッシュ関数として、ライベスト (R. Rivest)によって提案されたMD5ハッシュ関数を使用する。MD5ハッシュ関数のアルゴリズムについては、ネットワークワーキンググループ及びライベストによって開設されているウェブサイト2 (<http://www.kleinscmidt.com/edi/md5.htm>)に開示されている。また、そのMD5ハッシュ関数のアルゴリズムは、国際公開公報WO01/80431 A1にも開示されている。或るテキストデータ（テキストファイル）にそのMD5ハッシュ関数を施すことによって、128ビットの要約値が得られる。通常のコンピュータでも今後は処理単位が64ビットのCPUが使用されるようになると考えられるが、この場合に128 (=

2・64) ビットの要約値は非常に扱い易い長さである。この場合には、192 (= 3・64) ビットの要約値も比較的扱い易いと考えられる。

【0068】

また、本例では、そのMD5ハッシュ関数のプログラムとして、そのウェブサイト2において公開されている、RSAデータセキュリティー社(RSA Data Security Inc.)によって開発されたプログラムを使用した。

その要約値の使用方法の一例として、DNAの配列情報の供給者(情報処理装置10)は、所定の生物のDNAのヌクレオチドの配列を読み取り、これに対応するテキストデータより、上記のハッシュ関数を用いて要約値を算出し、この要約値をその生物の名称、及びDNAの位置を示す情報と共にインターネット上で閲覧可能にする。これによって、その供給者は、そのテキストデータ自体を公開することなく、その生物のDNAの配列情報を最先に解読したことを主張できると考えられる。その後、或るユーザからのその配列情報の購入希望が来たときに、その供給者は、そのヌクレオチドの配列のテキストデータを表1を用いてバイナリーデータに変換し、このバイナリーデータを例えば通信ネットワーク1を介して電子メールの形でそのユーザに送信する。これに対してユーザ側では、そのバイナリーデータを表1を用いてテキストデータに変換し、この逆変換されたテキストデータに上記のハッシュ関数を施して要約値を求める。

【0069】

そして、この要約値とその供給者によって公開されている要約値とが等しいときには、購入した配列情報が、供給の保持している配列情報と等しいことが極めて高い確率で保証される。更に、ユーザ側では、複数の供給者が公開している要約値を比較することによって、同じ配列情報を異なる複数の供給者から重複して購入することを防止することができる。これらの際に、ヌクレオチドの配列の長さ、及び先端部や末尾の一部の短い配列の比較を行うことによって、その配列情報の同一性を高めることができる。

【0070】

なお、ハッシュ関数としては、例えば文献2(FIPS Publication 180, 1993)で開示されているように、NBS(National Bureau of Standards)によって提案さ



れた S H S (Secure Hash Standard) ハッシュ関数を使用してもよい。S H S ハッシュ関数は、M D 5 ハッシュ関数よりも複雑な演算を行うと共に、1 6 0 ビットの要約値が得られる。これに関して、例えばタンパク質を構成するアミノ酸の配列数は 2 0 個～1 0 0 0 個程度であり、特に一文字表記を使用する際にはそれに対応するテキストデータも 2 0 バイト～1 k バイト程度に短くなるため、要約値から元のテキストデータが推定し易いと考えられる。そこで、アミノ酸の配列情報の要約値を求める際には、S H S ハッシュ関数を使用する方が望ましいことがある。

【0 0 7 1】

また、例えばヌクレオチドの配列を示す 2 つの膨大な長さのテキストデータの同一性を確認するために、ハッシュ関数の要約値を算出するような場合には、それ程複雑な計算を繰り返して行う必要は無いと考えられる。そこで、このような用途では、例えば文献 3 (R. L. Rivest: "The MD4 message digest algorithm", Lecture Notes in Computer Science, 537, 303-311 (1991)) で開示されている M D 4 ハッシュ関数を使用してもよいと考えられる。また、そのように単に同一性を確認する用途では、要約値の長さも 4 0 ビット～1 2 8 ビット程度でよい場合がある。

【0 0 7 2】

次に、本例の D N A 情報の供給者 (コンピュータシステム 2 A) と、ユーザ (コンピュータシステム 2 B) との間で D N A の配列情報を受け渡す際のビジネスモデルの一例につき図 3 ～図 6 のフローチャートを参照して詳細に説明する。先ず、D N A 情報の供給者側では、図 3 のステップ 1 0 1 において、シーケンサー 4 を使用して標準となる試料 (標準試料 E とする) の D N A 中の一方の系列のヌクレオチドの配列を読み取り、読み取った配列を表すテキストデータ T X 1 を情報処理装置 1 0 に供給する。本例では、その標準試料 E を大腸菌として、そのテキストデータ T X 1 として、図 7 に示すように、上記のウェブサイト 1 から入手した大腸菌の D N A の配列情報の内の、最初から 2 0 4 8 個までのヌクレオチドの配列を示すテキストデータを使用する。

【0 0 7 3】

標準試料 E の DNA 配列は配列番号 1 に示されている。図 7 のテキストデータは、配列番号 1 の配列から数字データを除いて、a, g, c, t の文字をそれぞれ A, G, C, T で置き換えたものに相当する。

次のステップ 1 0 2 において、情報処理装置 1 0 は、供給されたテキストデータ TX 1 に上記の MD 5 ハッシュ関数を施して 1 2 8 ビットの要約値 AB 1 を求めると共に、そのヌクレオチドの配列の数 NA 1、及び先頭と末尾との 8 個ずつのヌクレオチドの配列 ST 1, SB 1 を求める。テキストデータ TX 1 に対する具体的な値は下記の通りである。

【 0 0 7 4 】

AB 1 = hex(849339ac244cde42b5346ab5989aab61) ... ( 1 1 )

NA 1 = 2 0 4 8

ST 1 = AGCTTTTC, SB 1 = CGCGAAGG

次のステップ 1 0 3 において、情報処理装置 1 0 は、テキストデータ TX 1 を逆方向に並べ替えたテキストデータ TXR 1 (=GGAAGC...TTTCGA) を求め、このテキストデータ TXR 1 の MD 5 ハッシュ関数による要約値 ABR 1、及びこのテキストデータ TXR 1 の先頭と末尾との 8 個ずつのヌクレオチドの配列 STR 1, SBR 1 を求める。配列 STR 1, SBR 1 は、上記の配列 SB 1, ST 1 をそれぞれ逆方向に並べ替えることによって容易に求めることができる。これらの具体的な値は以下の通りである。

【 0 0 7 5 】

ABR 1 = hex(4eb1feae30f522642b912ce3ea09652b) ... ( 1 2 )

STR 1 = GGAAGCGC, SBR 1 = CTTTTTCGA

次のステップ 1 0 4 において、情報処理装置 1 0 は、標準試料 E の名前の情報（試料を特定する情報）、配列の数 NA 1、テキストデータ TX 1、配列 ST 1, SB 1、要約値 AB 1、逆方向の配列 STR 1, SBR 1、及び逆方向の要約値 ABR 1 を磁気ディスク装置 1 7 のマスターファイル 1 9 に記録する。この際に、マスターファイル 1 9 を複数のファイルとして、テキストデータ TX 1 と、それ以外のデータとを別のファイルに記録してもよい。また、テキストデータ TX 1 が例えば 1 0 0 M バイト程度以上になる場合には、テキストデータ TX 1 を

複数のマスターファイルに分割して記録してもよい。

#### 【0076】

次のステップ105において、情報処理装置10は、図7に示すように、標準試料EのテキストデータTX1を配列方向（ヌクレオチドの配列方向）にN行で、その配列方向に直交する方向（以下、「非配列方向」という）にM列の16文字の長さの部分テキストデータT(i, j) (i = 1 ~ N, j = 1 ~ M)に分割する。なお、N, Mはそれぞれ2以上の任意の整数であり、(4)式、(5)式を用いて既に説明したように、テキストデータTX1が100kバイト程度（又はこの整数倍）であるときに、このテキストデータTX1に対して1/20程度のデータ量のパリティ情報を得たい場合には、例えばNの値が64、Mの値が128に設定される。以下では説明を簡単にするために、図7に示すようにテキストデータTX1を4行で、かつ32列に分割した場合を想定する。即ち、N = 4, M = 32とする。この場合、本例では端数は生じないが、例えば図7において、最後の部分テキストデータT(4, 32)中の文字が16個より少ない場合には、足りない部分には予め定めた文字（例えば文字A）をダミーデータとして付加すればよい。また、部分テキストデータT(i, j)の長さは、16文字以外の任意の長さでよいが、処理速度を高めるためには、8文字の倍数が効率的である。

#### 【0077】

更に、情報処理装置10は、図7の16文字分の各部分テキストデータT(i, j)をそれぞれ所定の変換テーブルに従って128 (= 16 × 8) ビットのバイナリーデータ（数値データ）よりなる部分データA(i, j)に変換する。本例ではその変換テーブルとして、次のように部分テキストデータT(i, j)を単にアスキーコードに変換する関数asc(T(i, j))を用いる。

#### 【0078】

$$A(i, j) = \text{asc}(T(i, j)) \quad \dots (13)$$

なお、図7にT(3, 11)の変換例で示すように、関数asc(T(i, j))は、部分テキストデータT(i, j)の先頭の文字のコードが最下位桁となり、末尾の文字のコードが最上位桁となるように変換を行う。この際に、例えば

最後の列の部分データ  $A(i, 32)$  が 128 ビットにならないときには、その上位に予め定めた文字コード、又は数値データの 0 (hex(000...)) などのダミーデータが付加される。この結果、図 8 に示す 4 行で、32 列の部分データ  $A(i, j)$  が得られる。また、部分データ  $A(i, j)$  を対応するヌクレオチドの配列方向に連続して配列したときの集合体(数値データ)をバイナリーデータ  $BN1$  とする。図 7 の部分テキストデータ  $T(i, j)$  と図 8 の部分データ  $A(i, j)$  とは実質的に同じデータ量である。

【0079】

次に、本例では、部分データ  $A(i, j)$  をガロア体(Galois field)  $GF(2^m)$  上の元のベクトル表示とみなして、部分データ  $A(i, j)$  に対してガロア体  $GF(2^m)$  上の所定の演算を施す。本例の部分データ  $A(i, j)$  は 128 ビットであるため、 $m$  の値は 128 (64 の 2 倍) となり、ガロア体  $GF(2^{128})$  が使用される。また、本例ではガロア体  $GF(2^{128})$  上の既約多項式  $GF(X)$  及び生成元  $\alpha$  として次の式を使用する。なお、ガロア体  $GF(2^m)$  は、拡大ガロア体とも呼ばれることがある。

【0080】

$$GF(X) = 1 + X^{121} + X^{126} + X^{127} + X^{128} \quad \dots (14)$$

$$\alpha = X \quad \dots (15)$$

ガロア体  $GF(2^{128})$  上のベクトル表示では、 $GF(X)$  は hex(1110000100...01) となり、 $\alpha$  は hex(00...0010) となる。なお、生成元  $\alpha$  としては、 $(1+X)$  など也可以使用できる。また、ガロア体  $GF(2^{128})$  上の既約多項式としては、例えば次の既約多項式  $GF'(X)$  も使用でき、この既約多項式  $GF'(X)$  に対する生成元としては次の  $\alpha'$  などを使用できる。

【0081】

$$GF'(X) = 1 + X^{11} + X^{124} + X^{125} + X^{126} + X^{127} + X^{128} \quad \dots (14A)$$

$$\alpha' = 1 + X + X^2 \quad \dots (15A)$$

また、多項式  $GF(X)$  が既約であることは、例えば文献 4 (Van der Waerden, B. L. (1953), Modern Algebra(2 vols.), p.77, Ungar, New York) に記載してある「Kroneckerの方法」で確認することができる。また、 $GF(X)$  が既

約であることは、実用的には、ウェブサイト 3 (<http://archives.math.utk.edu/software/msdos/number.theory/ubasic/.html>)、又はウェブサイト 4 (<http://www.rkmath.rikkyo.ac.jp/~kida/ubasic.htm>) に開示されている整数論研究用のソフトウェアである「UBASIC」中の組み込み関数「POLFACT2」を用いても確認することができる。

【0082】

また、ガロア体  $GF(2^m)$  上の生成元  $\alpha$  は、 $k = 2^m - 1$  とおくと、既約多項式  $GF(X)$  を法として、次の関係を満たす。

$$\alpha^k = 1 \pmod{GF(X)} \quad \dots (16)$$

$$\alpha^{k'} \neq 1 \pmod{GF(X)} \quad (1 \leq k' < k) \quad \dots (17)$$

そこで、素数  $p_1, p_2, \dots, p_r$  及び整数  $n_1, n_2, \dots, n_r$  を用いて、 $k$  が次のように因数分解できるものとする。

【0083】

$$k = 2^m - 1 = p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r} \quad \dots (18)$$

このとき、生成元  $\alpha$  とは、既約多項式  $GF(X)$  を法として、 $\alpha$  の  $(p_1^{n_1-1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r})$  乗、 $(p_1^{n_1} \cdot p_2^{n_2-1} \cdot \dots \cdot p_r^{n_r})$  乗、 $\dots$ 、 $(p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_r^{n_r-1})$  乗が何れも 1 とならないものであればよい。

また、ガロア体  $GF(2^m)$  上の任意の 0 以外の元  $\beta$  についても (16) 式が成立するため、 $k (= 2^m - 1)$  を用いて  $\beta$  の逆元  $\beta^{-1}$  は次のように計算することも可能である。

【0084】

$$\beta^{-1} = \beta^{k-1} \pmod{GF(X)} \quad \dots (16R)$$

従って、例えば部分データ  $A(i, j)$  を  $\beta$  で除算する場合には、部分データ  $A(i, j)$  に  $\beta^{k-1}$  を乗算すればよい

次のステップ 106 において、情報処理装置 10 は、図 8 の各行 ( $i = 1 \sim 4$ ) の部分データ  $A(i, j)$  に対してガロア体  $GF(2^{128})$  上で、非配列方向 ( $j = 1 \sim 32$ ) に対する和である第 1 パリティ (Parity)  $B_1(i)$ 、 $\sum \alpha^{(j-1)} \cdot A(i, j)$  である第 2 パリティ  $B_2(i)$ 、及び  $\sum \alpha^{2(j-1)} \cdot A(i, j)$  である第 3 パリティ  $B_3(i)$  を計算する。これらの非配列方向のパリティ  $B_1$  (

$i) \sim B3(i)$  (第1組のパリティ情報) は、生成元  $\alpha$  を用いて、かつ既約多項式  $GF(X)$  を法として以下のように表すことができる。パリティ  $B1(i) \sim B3(i)$  における記号  $(\Sigma)$  は係数  $j$  に対する  $1 \sim 32$  の和を意味しており、以下の式は係数  $i$  の  $1 \sim 4$  の範囲で計算される。

【0085】

$$B1(i) = \Sigma A(i, j) = A(i, 1) + A(i, 2) + \dots + A(i, 32) \quad \dots (19)$$

$$B2(i) = \Sigma \alpha^{(j-1)} \cdot A(i, j) = A(i, 1) + \alpha \cdot A(i, 2) + \dots + \alpha^{31} \cdot A(i, 32) \quad \dots (20)$$

$$B3(i) = \Sigma \alpha^{2(j-1)} \cdot A(i, j) = A(i, 1) + \alpha^2 \cdot A(i, 2) + \dots + \alpha^{62} \cdot A(i, 32) \quad \dots (21)$$

この場合、(19) 式のパリティ  $B1(i)$  のベクトル表示は、部分データ  $A(i, j)$  についてビット毎に排他的論理和演算を行って得られる結果と同じである。また、(20) 式、(21) 式のパリティ  $B2(i)$ 、 $B3(i)$  は、それぞれ部分データ  $A(i, j)$  を (1) 式のように  $127$  次 ( $m=128$ ) 以下の多項式で表して、既約多項式  $GF(X)$  を法として演算を行うことによって計算することができる。

【0086】

更に、情報処理装置 10 は、図 8 の各列 ( $j=1 \sim 32$ ) の部分データ  $A(i, j)$  に対してガロア体  $GF(2^{128})$  上で、配列方向 ( $i=1 \sim 4$ ) に対する和である第1パリティ  $C1(j)$ 、 $\Sigma \alpha^{(i-1)} \cdot A(i, j)$  である第2パリティ  $C2(j)$ 、及び  $\Sigma \alpha^{2(i-1)} \cdot A(i, j)$  である第3パリティ  $C3(j)$  を計算する。これらの配列方向のパリティ  $C1(j) \sim C3(j)$  (第2組のパリティ情報) は、生成元  $\alpha$  を用いて、かつ既約多項式  $GF(X)$  を法として以下のように表すことができる。パリティ  $C1(j) \sim C3(j)$  における記号  $(\Sigma)$  は係数  $i$  に対する  $1 \sim 4$  の和を意味しており、以下の式は係数  $j$  の  $1 \sim 32$  の範囲で計算される。

【0087】

$$C1(j) = \Sigma A(i, j) = A(1, j)$$

$$+ A(2, j) + \dots + A(4, j) \quad \dots (22)$$

$$C2(j) = \sum \alpha^{(i-1)} \cdot A(i, j) = A(1, j)$$

$$+ \alpha \cdot A(2, j) + \dots + \alpha^3 \cdot A(4, j) \quad \dots (23)$$

$$C3(j) = \sum \alpha^{2(i-1)} \cdot A(i, j) = A(1, j)$$

$$+ \alpha^2 \cdot A(2, j) + \dots + \alpha^6 \cdot A(4, j) \quad \dots (24)$$

部分データ  $A(i, j)$  に対して実際にパリティ  $B1(i) \sim B3(i)$ 、及びパリティ  $C1(j) \sim C3(j)$  を計算した結果のベクトル表示が、図8に16進数表示で示されている。この例においては、各行のパリティ  $B1(i) \sim B3(i)$ 、及び各列のパリティ  $C1(j) \sim C3(j)$  はそれぞれ3個であるため、2つのヌクレオチドの配列の比較を行う場合に、各行及び各列において、それぞれ3個までの部分データ  $A(i, j)$  の相違部の復元を正確に行うことができる。従って、各行及び各列において、部分データ  $A(i, j)$  の相違部の位置の検出(特定)だけを行うと共に、相違部の復元を1個だけ行えばよい場合には、パリティ情報として、 $B1(i)$  及び  $C1(j)$ 、又は  $B2(i)$  及び  $C2(j)$  だけを使用(計算)するだけでもよい。後者のパリティ  $B2(i)$ 、 $C2(j)$  だけを使用する場合には、例えば或る行又は列で2つの部分データ  $A(i, j)$  が入れ替わったような配列であっても、配列の相違部の位置を検出できる利点がある。

#### 【0088】

また、各行及び各列において、それぞれ2個までの部分データ  $A(i, j)$  の相違部の復元を正確に行うことができればよい場合には、第1組のパリティ情報として  $B1(i)$ 、 $B2(i)$ 、 $B3(i)$  の何れか2つ、及び第2組のパリティ情報として  $C1(j)$ 、 $C2(j)$ 、 $C3(j)$  の何れか2つだけを使用(計算)するだけでもよい。また、各行と各列とで復元できる部分データの個数が違ってよい場合には、第1組のパリティ情報と第2組のパリティ情報とでパリティの個数が違ってよい。更に、各行又は各列において、それぞれ4個以上の相違部の復元を正確に行うためには、例えば  $\sum \alpha^{s(j-1)} \cdot A(i, j)$  であるパリティ  $B_s(i)$  ( $s=4, 5, \dots$ )、又は  $\sum \alpha^{t(i-1)} \cdot A(i, j)$  であるパリティ  $C_t(j)$  ( $t=4, 5, \dots$ ) を計算すればよい。

## 【0089】

また、図8の部分データA(i, j)の配列が実際には、64行×128列であるとする、図8の例のように、各行及び各列で3個までの相違部の復元を行う場合には、パリティB1(i)～B3(i), C1(j)～C3(j)はそれぞれ128ビット(16バイト)であるため、全部のパリティ情報のデータ量は、 $576 \cdot 16 (= (64 + 128) \cdot 3 \cdot 16)$  バイトとなる。一方、部分データA(i, j)のデータ量は、 $8192 \cdot 16 (= 64 \cdot 128 \cdot 16)$  バイトとなる。従って、全部のパリティ情報のデータ量は、全部の部分データ(i, j)に対してほぼ1/14程度に減少している。

## 【0090】

次に図4のステップ107において、情報処理装置10は、標準試料Eの名前の情報、配列の数NA1、バイナリーデータBN1、パリティB1(i)～B3(i), C1(j)～C3(j)を磁気ディスク装置17のワーキングファイル20に記録する。この際に、ワーキングファイル20を複数のファイルとして、バイナリーデータBN1と、パリティB1(i)～B3(i), C1(j)～C3(j)とを別のファイルに記録してもよい。更に、バイナリーデータBN1と共に、ステップ102で算出した要約値AB1をワーキングファイル20に記録してもよい。

## 【0091】

また、バイナリーデータBN1が長いときには、バイナリーデータBN1を複数のファイルに分割して記録してもよい。更に、図7のテキストデータTX1(ひいては図8のバイナリーデータBN1)がかなり長い場合には、テキストデータTX1を例えば数100kバイト程度を単位として複数のデータ群に分割し、各データ群毎にパリティB1(i)～B3(i), C1(j)～C3(j)を求めるようにしてもよい。

## 【0092】

更に、ステップ107において、DNA情報の供給者は、ワーキングファイル20に記録した情報、即ち標準試料Eの名前の情報、配列の数NA1、バイナリーデータBN1、パリティB1(i)～B3(i), C1(j)～C3(j)と



、マスターファイル19に記録した要約値AB1、ABR1の情報とを、CD-R/RWドライブ15を介してCD-R16に記録してもよい。このCD-R16から、更に多数のCD-ROMを作製してもよく、これらの記録媒体が郵送等によってユーザに販売される。

【0093】

次の、ステップ108において、情報処理装置10は、標準試料Eの名前の情報、配列の数NA1、配列ST1、SB1、要約値AB1、逆方向の配列STR1、SBR1、及び逆方向の要約値ABR1を磁気ディスク装置17のコンテンツファイル21に記録する。図7のテキストデータTX1が仮に100Mバイト程度の膨大なものであっても、コンテンツファイル21に記録されるデータは500バイト程度の僅かなものである。更に、情報処理装置10は、コンテンツファイル21中の情報を通信ネットワーク1を介してコンテンツのプロバイダ3に送信する。これによって、コンテンツファイル21中の情報はプロバイダ3のサーバ内の閲覧可能なコンテンツファイル31に記録されて、第3者がインターネットを介して自由に閲覧できるようになる。

【0094】

次のステップ109において、DNA情報の供給者は、ユーザから購入要求が来るのを待つ状態となる。そして、(a)ユーザから標準試料Eに対する簡易データの要求があったときには、ステップ110に移行して、情報処理装置10は、磁気ディスク装置17のワーキングファイル20の中のパリティ情報(パリティB1(i)～B3(i)、C1(j)～C3(j))を例えば電子メールの添付ファイルとしてそのユーザに送信する。一方、ステップ109において、(b)ユーザから完全データの要求があったときには、ステップ111に移行して、情報処理装置10は、ワーキングファイル20中のバイナリーデータBN1をZIPファイル等の形式で圧縮し、この圧縮されたデータを例えば電子メールの添付ファイルとしてそのユーザに送信する。この際に必要に応じて、ハッシュ関数による要約値AB1を同時に送信してもよい。本例によれば、簡易データ(パリティ情報)はデータ量が少ないために短時間で送信することができる。

【0095】

また、ステップ109において、ユーザは、必要に応じて部分データ、即ち図8の全部の部分データA(i, j)の内の所望のデータ、例えば2つの部分データA(4, 16)及びA(1, 17)のみをその供給者から購入するようにしてもよい。これによって、必要な正確なデータのみを短時間に入手することができる。

#### 【0096】

次に、DNA情報のユーザ（図1のコンピュータシステム2Bの所有者とする）側では、図5のステップ121において、図1の通信ネットワーク1（インターネット）を介してプロバイダ3のサーバ内のコンテンツファイル31の内容を閲覧し、その中からステップ108で供給者（図1のコンピュータシステム2A）から送信された情報、即ち標準試料Eの名前の情報、ヌクレオチドの配列の数NA1、配列ST1, SB1、要約値AB1、逆方向の配列STR1, SBR1、及び逆方向の要約値ABR1を読み取り、読み取った情報をコンピュータシステム2B内の記憶装置の一時ファイルに記録する。

#### 【0097】

次の、ステップ122において、そのユーザは、不図示のDNAのシーケンサーを用いて、標準試料Eと同じ種類で検査対象の試料FのDNA中の一方の系列のヌクレオチドの配列を読み取り、読み取られた配列を示すテキストデータTX2（アスキーコードとする）をコンピュータシステム2B内の情報処理装置に取り込む。その検査対象の試料Fとは、例えば突然変異を起こしていると思われる大腸菌であり、そのテキストデータTX2は、標準試料EのテキストデータTX1と同様に最初から2048個までのヌクレオチドの配列を示すものとする。

#### 【0098】

試料FのDNA配列は配列番号2に示されている。後述の図9のテキストデータは、配列番号2の配列から数字データを除いて、a, g, c, tの文字をそれぞれA, G, C, Tで置き換えたものに相当する。

図9は、その試料FのDNAのヌクレオチドの配列に対応するテキストデータTX2を示し、この図9の配列の内のアンダーラインを付した部分のみが、図7の標準試料Eの配列と異なっている。即ち、試料Fの配列は、標準試料Eの部分

テキストデータ T (4, 16), T (1, 17) の部分だけが以下のように異なっている。なお、この段階では、ユーザは、試料 F の配列と標準試料 E の配列とのどの部分が相違しているのかは分からない。

【 0 0 9 9 】

標準試料 E

試料 F

T(4,16)=ATTTGGACGGACGTTG → ATTTGGACATTATGGC

T(1,17)=ACGGGGTCTATACCTG → GGCCAACTTATACCTG

そして、ユーザのコンピュータシステム 2 B 側の情報処理装置においても、DNA の配列情報を処理するためのアプリケーション・プログラムが起動されている。そして、その情報処理装置は、ステップ 1 2 3 において、読み取られたテキストデータ TX 2 に上記の MD 5 ハッシュ関数を施して 1 2 8 ビットの要約値 AB 2 を求めると共に、そのヌクレオチドの配列の数 NA 2、及び先頭と末尾との 8 個ずつのヌクレオチドの配列 ST 2, SB 2 を求め、これらを内部の記憶装置の第 1 データファイルに記録する。テキストデータ TX 2 (図 9) に対する具体的な値は下記の通りである。

【 0 1 0 0 】

AB 2 = hex(1457b51222a83c3222e87cb4d4e63305) ... (2 5)

NA 2 = 2 0 4 8

ST 2 = AGCTTTTC, SB 2 = CGCGAAGG

次のステップ 1 2 4 において、情報処理装置は、試料 F の配列数 NA 2 と標準試料 E の配列数 NA 1 とが等しいかどうかを調べ、両者が異なっている場合には、ユーザはステップ 1 2 5 に移行して、別の DNA 情報を検索し、NA 2 と同じ配列数の DNA 情報をサーチする。本例では、ステップ 1 2 4 において、NA 2 = NA 1 であるため、動作はステップ 1 2 6 に移行して、試料 F の先頭と末尾との一部の配列 ST 2, SB 2 が、標準試料 E の配列 ST 1, SB 1 と等しいかどうか、更に試料 F の要約値 AB 2 が標準試料 E の要約値 AB 1 (ステップ 1 2 1 で一時ファイルに記録されている) と等しいかどうかを調べる。これらが共に等しい場合には、試料 F の配列と標準試料 E の配列とは非常に高い確率 (ほぼ  $1 / 2^{128} \doteq 1 / 10^{38}$  程度の確率) で一致しているとみなすことができる。従って

、ステップ127に移行して、コンピュータシステム2Bの情報処理装置は、その第1データファイルに「試料FのDNA構造は、標準試料EのDNA構造と同一である。」との情報を記録する。

## 【0101】

但し、本例では、 $ST2 = ST1$ 、 $SB2 = SB1$ が成立するが、(11)式及び(25)式より $AB2 \neq AB1$ であるため、動作はステップ126からステップ128に移行して、その情報処理装置は、試料Fの先頭と末尾との一部の配列 $ST2$ 、 $SB2$ が、標準試料Eを逆に並べた配列の一部の配列 $STR1$ 、 $SBR1$ と等しいかどうか、更に試料Fの要約値 $AB2$ が標準試料Eを逆に並べた配列の要約値 $ABR1$ と等しいかどうかを調べる。これらが共に等しい場合には、試料Fの配列と標準試料Eを逆に並べた配列とは非常に高い確率で一致しているとみなすことができる。従って、ステップ139に移行して、コンピュータシステム2Bの情報処理装置は、その第1データファイルに「試料FのDNA構造は、標準試料EのDNA構造に対して回文 (palindrome) の関係にある。」との情報を記録する。

## 【0102】

本例では、 $ST2 \neq STR2$ 、 $SB2 \neq SBR2$ 、かつ(12)式及び(25)式より $AB2 \neq ABR1$ であるため、動作はステップ128からステップ129に移行して、そのユーザは、通信ネットワーク1 (インターネット) を介してDNA情報の供給者から上記の簡易データ、即ち標準試料Eのパリティ情報 ( $B1(i) \sim B3(i)$ 、 $C1(j) \sim C3(j)$ ) (図8の情報) を購入し、購入した情報をコンピュータシステム2B (情報処理装置) 内の記憶装置の第2データファイルに記録する。

## 【0103】

次に、図6のステップ130において、コンピュータシステム2Bの情報処理装置は、図9に示すように、試料Fのテキストデータ $TX2$ を配列方向 (ヌクレオチドの配列方向) にN行で、非配列方向にM列の16文字の長さの部分テキストデータ $TF(i, j)$  ( $i = 1 \sim N$ 、 $j = 1 \sim M$ ) に分割する。分割数N、Mは標準試料Eの分割数と同じであり、本例では、 $N = 4$ 、 $M = 32$ である。更に

、情報処理装置は、図9の各部分テキストデータ $TF(i, j)$ を次のようにテキストデータを単にアスキーコードに変換する関数 $asc(TF(i, j))$ を用いて、128(=16・8)ビットのバイナリーデータ(数値データ)よりなる部分データ $AF(i, j)$ に変換する。この場合にも、部分テキストデータ $TF(i, j)$ の文字列は反転してアスキーコード列に変換される。

## 【0104】

$$AF(i, j) = asc(TF(i, j)) \quad \cdots (26)$$

この結果、図10に示す4行で、32列の部分データ $AF(i, j)$ が得られる。また、部分データ $AF(i, j)$ を連続して配列した集合体(数値データ)をバイナリーデータ $BN2$ とする。

次に、情報処理装置は、ステップ106の動作と同様にして、図10の各行( $i=1\sim4$ )の部分データ $AF(i, j)$ に対してガロア体 $GF(2^{128})$ 上で、非配列方向( $j=1\sim32$ )に対する和である第1パリティ(Parity) $B1F(i)$ 、 $\sum \alpha^{(j-1)} \cdot AF(i, j)$ である第2パリティ $B2F(i)$ 、及び $\sum \alpha^{2(j-1)} \cdot AF(i, j)$ である第3パリティ $B3F(i)$ を計算する。これらの非配列方向のパリティ $B1F(i) \sim B3F(i)$ (第1組のパリティ情報)は、(15)式の生成元 $\alpha$ を用いて、かつ(14)式の既約多項式 $GF(X)$ を法として(19)式～(21)式と同様に、係数 $i$ について1～4の範囲で計算される。

## 【0105】

次に、その情報処理装置は、図10の各列( $j=1\sim32$ )の部分データ $AF(i, j)$ に対してガロア体 $GF(2^{128})$ 上で、配列方向( $i=1\sim4$ )に対する和である第1パリティ $C1F(j)$ 、 $\sum \alpha^{(i-1)} \cdot AF(i, j)$ である第2パリティ $C2F(j)$ 、及び $\sum \alpha^{2(i-1)} \cdot AF(i, j)$ である第3パリティ $C3F(j)$ を計算する。これらの配列方向のパリティ $C1(j) \sim C3(j)$ (第2組のパリティ情報)も、(15)式の生成元 $\alpha$ を用いて、かつ(14)式の既約多項式 $GF(X)$ を法として(22)式～(24)式と同様に、係数 $j$ について1～32の範囲で計算される。

## 【0106】

部分データ  $AF(i, j)$  に対して実際にパリティ  $B1F(i) \sim B3F(i)$ 、及びパリティ  $C1F(j) \sim C3F(j)$  を計算した結果のベクトル表示が、図10に16進数表示で示されている。

次に、ステップ131において、その情報処理装置は、供給者から購入した簡易データの2組のパリティ、即ち図8（標準試料E）の2組のパリティ  $B1(i) \sim B3(i)$ 、 $C1(j) \sim C3(j)$  と、図10（試料F）の2組のパリティ  $B1F(i) \sim B3F(i)$ 、 $C1F(j) \sim C3F(j)$  とを比較して、相違するパリティをサーチする。本例では、図8（標準試料F）に対して図10（試料F）の  $i=1, 4$  の非配列方向のパリティ  $B1F(1) \sim B3F(1)$ 、 $B1F(4) \sim B3F(4)$  と、 $j=16, 17$  の配列方向のパリティ  $C1F(16) \sim C3F(16)$ 、 $C1F(17) \sim C3F(17)$  とが異なっている。なお、各行のパリティ  $B1F(i) \sim B3F(i)$ 、又は各列のパリティ  $C1F(j) \sim C3F(j)$  において、1つでもパリティが異なっていれば、その行又は列のパリティが異なっているとみなすことができる。

#### 【0107】

従って、図10（試料F）の部分データ  $AF(i, j)$  において、 $i=1, 4$  の行と  $j=16, 17$  の列との交点に位置する4つの部分データ  $AF(1, 16)$ 、 $AF(4, 16)$ 、 $AF(1, 17)$ 、 $AF(4, 17)$  が図8（標準試料E）と相違すると特定できる。また、これ以外の試料Fの部分データ  $AF(i, j)$  は標準試料Eの部分データ  $A(i, j)$  とほぼ同一であるとみなすことができる。

#### 【0108】

また、図11は、主に図10の試料Fのデータ中から図8と異なるパリティ  $B1F(1) \sim B3F(1)$ 、 $B1F(4) \sim B3F(4)$ 、 $C1F(16) \sim C3F(16)$ 、 $C1F(17) \sim C3F(17)$  を取り出して表示したものである。また、図11において図8と異なる部分データ  $AF(1, 16)$ 、 $AF(4, 16)$ 、 $AF(1, 17)$ 、 $AF(4, 17)$  の位置に、復元すべきデータ  $X1$ 、 $X2$ 、 $Y1$ 、 $Y2$  を表示している。この復元すべきデータ  $X1$ 、 $X2$ 、 $Y1$ 、 $Y2$  はそれぞれ図8（標準試料E）の部分データ  $A(1, 16)$ 、 $A(4, 1$

6),  $A(1, 17)$ ,  $A(4, 17)$  である。

【0109】

次のステップ132において、その情報処理装置は、図10の部分データ  $AF(i, j)$  中で図8の部分データ  $A(i, j)$  と相違する部分データ ( $AF(i', j')$  とする) は、各行、又は各列に多くとも3つかどうかを調べる。これが成立する場合には、その部分データ  $AF(i', j')$  に対応する標準試料Eの部分データ  $A(i', j')$  は、ガロア体  $GF(2^{128})$  上で連立方程式を解くことによって正確に求める(復元する)ことができる。本例では、それが成立する、即ち相違する変換データは、第1行、第4行に2つずつで、かつ第16列、第17列に2つずつであるため、動作はステップ133に移行する。そして、その情報処理装置は、2組の相違するパリティ、及び試料Fの相違する部分データ  $AF(i', j')$  を用いて、図12のフローチャートに従って対応する標準試料Eの部分データ  $A(i', j')$  ( $X1, X2, Y1, Y2$ ) を復元する。図12の計算は、全てガロア体  $GF(2^{128})$  上で実行される。

【0110】

この場合、図11において、第16列の未知数  $X1, X2$  は2つであるため、第16列の2つのパリティ  $C1F(16)$ ,  $C2F(16)$  と、対応する図8の2つのパリティ  $C1(16)$ ,  $C2(16)$  と、未知数  $X1, X2$  に対応する試料Fの部分データ  $AF(1, 16)$ ,  $AF(4, 16)$  とを用いて2元1次連立方程式を組み立てる。即ち、パリティ  $C1(16)$ ,  $C1F(16)$  に対する計算式は図12のステップ141の(G1)式、(G2)式となる。また、(15)式の生成元  $\alpha$  を用いて、パリティ  $C2(16)$ ,  $C2F(16)$  に対する計算式はステップ142の(G3)式、(G4)式となる。

【0111】

次に、(G1)式から(G2)式を引き、(G3)式から(G4)式を引くことで、それぞれステップ143の(G5)式、(G6)式が得られる。(G5)式、(G6)式の右辺をそれぞれ  $C1X$ ,  $C2X$  とすることで、2元1次連立方程式が得られる。そこで、これを解くことによって、未知数  $X1, X2$  はそれぞれステップ144の(G7)式で表すことができる。これを実際に解いた結果、

X1, X2は次のようになる(図11参照)。なお、未知数が3個であれば、第3パリティC3(16), C3F(16)なども用いて、3元1次連立方程式を解けばよく、未知数が1個であれば、例えば第1パリティC1(16), C1F(16)などを用いるだけでよい。

【0112】

$$X1 = \text{hex}(43475447474347544347544354434154) \quad \dots (27)$$

$$X2 = \text{hex}(47545447434147474341474754545441) \quad \dots (28)$$

更に、アスキーコード列を文字列に変換する関数chr()を用いて、この数値データを文字列に変換すると次のようになる(図11参照)。この関数chr()は、上記の関数asc()と対称に、アスキーコード列を1バイト単位で最大桁のコードが末尾の文字となり、最小桁のコードが先頭の文字になるように反転して文字列に変換する。

【0113】

$$\text{chr}(X1) = \text{TACTCTGCTGCGGTGC}$$

$$= T(1, 16) = TF(1, 16) \quad \dots (29)$$

$$\text{chr}(X2) = \text{ATTTGGACGGACGTTG} = T(4, 16) \quad \dots (30)$$

これより、標準試料Eの部分テキストデータT(1, 16)と試料Fの部分テキストデータTF(1, 16)とは等しく、部分テキストデータT(4, 16)だけが部分テキストデータTF(4, 16)(図9参照)と異なることが分かる。

【0114】

次に、図11において、第17列の未知数Y1, Y2についても、第17列の2つのパリティC1F(17), C2F(17)と、対応する図8の2つのパリティC1(17), C2(17)と、未知数Y1, Y2に対応する試料Fの部分データAF(1, 17), AF(4, 17)とを用いて、図12のステップ145の(G8)式、(G9)式よりなる2元1次連立方程式が得られる。これを解くことによって、未知数Y1, Y2はそれぞれステップ146の(G10)式で表すことができる。これを実際に解いた結果、Y1, Y2は次のようになる(図11参照)。



【 0 1 1 5 】

$Y\ 1 = \text{hex}(47544343415441544354474747474341) \quad \dots (3\ 1)$

$Y\ 2 = \text{hex}(41544343544754414743544741414754) \quad \dots (3\ 2)$

更に、この数値データ（アスキーコード列）を文字列に変換すると次のようになる（図 1 1 参照）。

$\text{chr}(Y\ 1) = \text{ACGGGGTCTATACCTG} = T(1, 17) \quad \dots (3\ 3)$

$\text{chr}(Y\ 2) = \text{TGAAGTCGATGTCCTA}$

$= T(4, 17) = TF(4, 17) \quad \dots (3\ 4)$

これより、標準試料 E の部分テキストデータ  $T(4, 17)$  と試料 F の部分テキストデータ  $TF(4, 17)$  とは等しく、部分テキストデータ  $T(1, 17)$  だけが部分テキストデータ  $TF(1, 17)$ （図 9 参照）と異なることが分かる。また、本例の方法によって未知数  $X\ 1$ ,  $X\ 2$ ,  $Y\ 1$ ,  $Y\ 2$ 、即ち標準試料 E の部分データ  $A(1, 16)$ ,  $A(4, 16)$ ,  $A(1, 17)$ ,  $A(4, 17)$  が正確に復元できていることが分かる。なお、部分データ  $A(1, 16)$ ,  $A(4, 17)$  は、それぞれ部分データ  $AF(1, 16)$ ,  $AF(4, 17)$  と同一であるため、復元されたデータとみなす必要はない。

【 0 1 1 6 】

次のステップ 1 3 4 において、その情報処理装置は、復元された部分データ  $A(i', j')$ 、即ち  $A(4, 16)$ ,  $A(1, 17)$  で、図 1 0 の試料 F のバイナリーデータ  $BN\ 2$  中の対応する部分データ  $AF(4, 16)$ ,  $AF(1, 17)$  を置き換えた後、この置き換えによって得られるバイナリーデータ  $BN\ 2$  をテキストデータ  $TX\ 1'$  に逆変換する。更に情報処理装置は、そのテキストデータ  $TX\ 1'$  より MD 5 ハッシュ関数を用いて 1 2 8 ビットの要約値  $AB\ 1'$  を算出し、この要約値  $AB\ 1'$  が標準試料 E の要約値  $AB\ 1$ （ステップ 1 2 1 で一時ファイルに記録されている）と等しいかどうかを確認する。本例では、 $AB\ 1' = AB\ 1$  が成立するが、例えば図 1 0 の試料 F の部分データ  $AF(i, j)$  中の相違する部分のデータの状態によって、その相違がパリティ情報に反映されないような場合には、その相違する部分の位置がステップ 1 3 2 で正確に検出されない可能性がある。このような場合に、 $AB\ 1' \neq AB\ 1$  となったときには、ステ

ップ135に移行すればよい。通常は、 $AB1' = AB1$ が成立して、動作はステップ138に移行して、情報処理装置は、上記の第1データファイルに「試料Fの配列と標準試料Eの配列との内で相違する部分の位置( $i'$ ,  $j'$ )、及び相違する部分テキストデータの対」の情報を記録する。本例では、位置( $i'$ ,  $j'$ )として位置(4, 16), (1, 17)が、相違する部分テキストデータの対としてA(4, 16), AF(4, 16)及びA(1, 17), AF(1, 17)が記録される。

【0117】

一方、ステップ132において、相違する部分データAF( $i'$ ,  $j'$ )の個数が4個以上の行、又は列が存在する場合には、その行又は列での部分データの正確な復元は困難である。そこで、動作はステップ135に移行して、そのユーザはそのDNA情報の供給者から標準試料Eの完全データ、即ち図8のバイナリーデータBN1を通信ネットワーク1(インターネット)を介して購入し、コンピュータシステム2Bの情報処理装置は、そのバイナリーデータBN1を記憶装置の第3データファイルに記録する。

【0118】

次のステップ136において、その情報処理装置は、そのバイナリーデータBN1をテキストデータTX1'に逆変換し、そのテキストデータTX1'よりMD5ハッシュ関数を用いて128ビットの要約値AB1'を算出し、この要約値AB1'が標準試料Eの要約値AB1(ステップ121で一時ファイルに記録されている)と等しいかどうかを確認する。通常は、 $AB1' = AB1$ が成立するが、例えば通信エラー等によって送信されたバイナリーデータBN1の中にエラーが生じている場合には、 $AB1' \neq AB1$ となる。このときには、例えば供給者に完全データの再送信を要請する等の対処を行う。そして、ステップ136で $AB1' = AB1$ が成立するときには、ステップ137に移行して情報処理装置は、標準試料EのバイナリーデータBN1中で、試料Fの相違している部分データAF( $i'$ ,  $j'$ )に対応する部分データA( $i'$ ,  $j'$ )を求める。その後、動作はステップ138に移行する。

【0119】

なお、上記のステップ135では、ユーザはDNA情報の供給者から完全データ（バイナリーデータBN1）を購入しているが、別の方法として、ステップ131で特定された相違する部分データAF（i'，j'）に対応する標準試料Eの部分データA（i'，j'）のみを購入してもよい。これによって、通信コストを大幅に低減できる。

#### 【0120】

このように本例のビジネスモデルによれば、第1段階として標準試料Eのパリティ情報（B1（i）～B3（i），C1（j）～C3（j））を購入している。次に、このパリティ情報と試料Fのパリティ情報（B1F（i）～B3F（i），C1F（j）～C3F（j））とを比較して、相違する部分データAF（i，j）の個数が少ない場合には、対応する標準試料Eの部分データA（i，j）を復元することとして、相違する部分データの個数が多い場合に、完全データ、又は相違する部分データのみを購入している。従って、初めから膨大な完全データを購入する必要がなく、通信時間を短縮できると共に、情報処理コストを低減できる。

#### 【0121】

また、本例のパリティ情報を用いれば、SNP（一塩基変位多型：Single Nucleotide Polymorphism）のように所定の範囲内で1つのヌクレオチド（塩基）だけが異なっているような異常は、容易にその位置の検出、及び復元を行うことができる。

なお、上記の実施の形態では、DNA情報のユーザは、ステップ121において、コンテンツファイルより標準試料Eの配列ST1，SB1、要約値AB1、及び配列STR1，SBR1、要約値ABR1を読み取って、ステップ122～128において、標準試料Eと試料Eとの同一性の判定を行って、両者が異なる場合に標準試料Eのパリティ情報（簡易データ）を購入している。しかしながら、標準試料Eと試料Fとは通常はいくらかは異なっていると考えられるため、このような要約値AB1等の読み取りから2つの試料の同一性の判定までの動作を省略して、すぐにステップ129に移行して、DNA情報の供給者から標準試料Eのパリティ情報（簡易データ）を購入するようにしてもよい。

## 【0122】

次に、上記の実施の形態では、ステップ105において、図7の標準試料EのテキストデータTX1を図8の同じデータ量の部分データA(i, j)の配列に変換し、この配列からパリティ情報を求めている。その代わりに、データ量を減少させるために、図7の標準試料EのテキストデータTX1を表1の変換テーブル(1つのヌクレオチドを2ビットのデータで表すテーブル)を用いてデータ量が1/4のバイナリーデータ(数値データ)に変換し、このバイナリーデータを配列方向、及び非配列方向に分割して、部分データの配列を作成してもよい。

## 【0123】

図13は、そのようにして得られたヌクレオチドの配列方向に5行で、非配列方向に13列の64ビット(8バイト)ずつの部分データB(i, j)(i=1~5, j=1~13)の配列を16進数表示で示し、この図13の各部分データB(i, j)は、それぞれ図7の標準試料E中の32個のヌクレオチドの配列に対応している。なお、この配列では、最後の部分データB(5, 13)に対応する図7の標準試料Eは存在しないため、その部分データB(5, 13)にはhex(000...000)よりなるダミーデータが付加されている。

## 【0124】

この場合には、ステップ106に対応して、図13の64ビットの各部分データB(i, j)をガロア体GF(2<sup>64</sup>)(m=64)上の元のベクトル表示とみなして、部分データB(i, j)に対してガロア体GF(2<sup>64</sup>)上の所定の演算を施す。ガロア体GF(2<sup>64</sup>)上の既約多項式GF(X)及び生成元αとしては次の式を使用することができる。

## 【0125】

$$GF(X) = 1 + X^5 + X^{23} + X^{43} + X^{64} \quad \dots (35)$$

$$\alpha = X \quad \dots (36)$$

また、ガロア体GF(2<sup>64</sup>)上の既約多項式としては、例えば次の既約多項式GF'(X)も使用でき、この既約多項式GF'(X)に対する生成元としては次のα'などを使用できる。

## 【0126】

$$GF'(X) = 1 + X^7 + X^{62} + X^{63} + X^{64} \quad \dots (35A)$$

$$\alpha' = 1 + X \quad \dots (36A)$$

そして、図1の情報処理装置10は、図13の各行 ( $i = 1 \sim 5$ ) の部分データ  $B(i, j)$  に対してガロア体  $GF(2^{64})$  上で、非配列方向 ( $j = 1 \sim 13$ ) に対する和である第1パリティ (Parity)  $B1B(i)$ 、 $\sum \alpha^{(j-1)} \cdot B(i, j)$  である第2パリティ  $B2B(i)$ 、及び  $\sum \alpha^{2(j-1)} \cdot B(i, j)$  である第3パリティ  $B3B(i)$  を計算する。これらの計算式は、(19)式～(21)式に対応しており、この第1組のパリティ情報 ( $B1B(i) \sim B3B(i)$ ) はそれぞれ64ビットである。

#### 【0127】

更に、情報処理装置10は、図13の各列 ( $j = 1 \sim 13$ ) の部分データ  $B(i, j)$  に対してガロア体  $GF(2^{64})$  上で、配列方向 ( $i = 1 \sim 5$ ) に対する和である第1パリティ  $C1B(j)$ 、 $\sum \alpha^{(i-1)} \cdot B(i, j)$  である第2パリティ  $C2B(j)$ 、及び  $\sum \alpha^{2(i-1)} \cdot B(i, j)$  である第3パリティ  $C3B(j)$  を計算する。これらの計算式は、(22)式～(24)式に対応しており、この第2組のパリティ情報 ( $C1B(j) \sim C3B(j)$ ) もそれぞれ64ビットである。

#### 【0128】

この場合、2組のパリティ情報 ( $B1B(i) \sim B3B(i)$ 、 $C1B(j) \sim C3B(j)$ ) のデータ量は、図8のパリティ情報 ( $B1(i) \sim B3(i)$ 、 $C1(j) \sim C3(j)$ ) に比べてほぼ1/4に少なくできる。従って、パリティ情報を通信回線を介して更に短時間で送信することができると共に、記録媒体に記録する場合にも、低容量の記録媒体を使用できる。

#### 【0129】

この場合には、ユーザ側で図9の試料Fのパリティ情報を計算する場合にも、同様にその試料Fのテキストデータを表1に従って1/4のデータ量の部分データの配列に変換した後、ガロア体  $GF(2^{64})$  上の演算によって第1組及び第2組のパリティ情報を計算すればよい。この後の相違するデータの位置の特定、及び元のデータの復元は上記の実施の形態と同様に行うことができる。

## 【0130】

なお、上記の実施の形態では、DNA又はRNAを構成するヌクレオチドは4種類であるため、テキストデータTX1を少ないデータ量のバイナリーデータに変換する際に、表1に示すように各ヌクレオチドを2ビットのデータで表している。これに対して、ヌクレオチド（又は塩基）を表すテキストデータとして、以下のような16種類の文字a～n（8ビットのアスキーデータ）が使用されることがある。

## 【0131】

- a アデニン（アデニンを含むヌクレオチドと同義、以下同様）
- c シトシン
- g グアニン
- t チミン
- u ウラシル
- m アデニン、又はシトシン
- r グアニン、又はアデニン
- w アデニン、又はチミン（若しくはウラシル）
- s グアニン、又はシトシン
- y チミン（若しくはウラシル）、又はシトシン
- k グアニン、又はチミン（若しくはウラシル）
- v アデニン、グアニン、又はシトシン
- h アデニン、シトシン、又はチミン（若しくはウラシル）
- d アデニン、グアニン、又はチミン（若しくはウラシル）
- b グアニン、シトシン、又はチミン（若しくはウラシル）
- n （アデニン、シトシン、グアニン、又はチミン（若しくはウラシル））

又は（不明若しくは他の塩基）。

## 【0132】

この場合には、これら16種類の文字を互いに異なる4ビットのコードに変換し、このコードを用いてテキストデータを数値データ（バイナリーデータ）に変換してもよい。これによって、データ量を1/2にすることができる。また、将

来的にヌクレオチド（塩基）の種類が増加したような場合には、これらのヌクレオチドを5ビット、又は6ビットのデータで表現するようにしてもよい。

## 【0133】

また、上記の実施の形態では、図7及び図9のヌクレオチドの配列を示すテキストデータよりハッシュ関数によって要約値を算出しているが、情報量としては、それらのテキストデータは例えば表1に従って変換したバイナリーデータ（数値データ）と等価である。従って、これらの変換後のバイナリーデータよりハッシュ関数によってそれぞれ要約値を算出し、これらの算出結果同士を比較するようにしてもよい。そのバイナリーデータのデータ量はテキストデータに対して1/4程度であるため、要約値を算出する時間が短縮できる利点がある。

## 【0134】

なお、上記の実施の形態では、DNA又はRNA中のヌクレオチドの配列（又は塩基の配列）の情報を処理対象としているが、本発明は、遺伝子を形成するヌクレオチドの配列の情報を処理する場合にも適用できることは言うまでもない。

次に、本発明の実施の形態の他の例につき説明する。本例は、タンパク質又はペプチドを構成するアミノ酸（生物学的物質）の配列情報を処理する場合に本発明を適用したものである。

## 【0135】

本例でも基本的に図1のコンピュータシステム2Aを使用できるが、DNAのシーケンサー4の代わりに、タンパク質のアミノ酸の配列を決定する配列読み取り装置としてのタンパク質用のシーケンサー（protein Sequencer）が情報処理装置10に接続される点が異なっている。なお、その配列読み取り装置としては、アミノ酸の配列のデータベースも使用できる。本例でも、例えば新規の試料Gのタンパク質のアミノ酸の配列をそのシーケンサーで解読した場合に、その配列を示すテキストデータ（TX3とする）が情報処理装置10に供給される。本例では一文字表記を採用するものとして、n個のアミノ酸の配列に対応するテキストデータ（アスキーコードとする）は、nバイトの長さである。本例では、その試料Gを大腸菌として、そのテキストデータTX3として、図14に示すように、上記のウェブサイト1から入手した大腸菌の或るタンパク質の820個のアミノ

酸の配列を示すテキストデータを使用する。

【0136】

試料Gのアミノ酸配列は配列番号3に示されている。図14のテキストデータは、配列番号3の配列から数字データを除いて、その配列を一文字表記で表したものに相当する。また、図14においては、そのテキストデータが配列方向（アミノ酸の配列方向）に4行で、その配列方向に直交する非配列方向に26列の8文字の長さの部分テキストデータTG(i, j)に分割されており、821番以上のアミノ酸を示すデータの位置にはダミーデータとして0が付加されている。

【0137】

次に、情報処理装置10は、供給されたテキストデータTX3に上記のMD5ハッシュ関数を施して128ビットの要約値AB3を求めると共に、そのアミノ酸の配列の数NA3、及び先頭と末尾との8個ずつのアミノ酸の配列ST3, SB3を求める。テキストデータTX3に対する具体的な値は下記の通りである。

AB3 = hex(0f66dc2b3024a9739d0e912fde12b8ba) ... (41)

NA3 = 820

ST3 = MRVLKFGG, SB3 = TLSWKLG V

次に、情報処理装置10は、テキストデータTX3を逆方向に並べ替えたテキストデータTXR3 (= VGLKWS ... FKLVRM)を求め、このテキストデータTXR3のMD5ハッシュ関数による要約値ABR3、及びこのテキストデータTXR3の先頭と末尾との8個ずつのアミノ酸の配列STR3, SBR3を求める。配列STR3, SBR3は、上記の配列SB3, ST3をそれぞれ逆方向に並べ替えることによって容易に求めることができる。これらの具体的な値は以下の通りである。テキストデータTXR3の配列は、テキストデータTX3の配列に対して回文 (palindrome) の関係にあるとすることができる。

【0138】

ABR3 = hex(e895f433e1e77f84b3cadeead1a52380) ... (42)

STR3 = VGLKWSLT, SBR3 = GGFKLVRM

次に、情報処理装置10は、試料Gの名前の情報（試料を特定する情報）、配列の数NA3、テキストデータTX3、配列ST3, SB3、要約値AB3、逆



方向の配列STR3, SBR3、及び逆方向の要約値ABR3を磁気ディスク装置17のマスターファイル19に記録する。この際に、マスターファイル19を複数のファイルとして、テキストデータTX3と、それ以外のデータとを別のファイルに記録してもよい。続いて、情報処理装置10は、例えば図7と同様に図14に示すように、試料GのテキストデータTX3を配列方向（アミノ酸の配列方向）にN行で、その配列方向に直交する非配列方向にM列の8文字（64ビット）の長さの部分テキストデータTG(i, j)に分割する。N, Mはそれぞれ2以上の任意の整数である。本例ではテキストデータTX3に例えば12文字分のダミーデータ（本例では0であるが、それ以外に例えば文字Aなども使用できる）を付加して得られる832(=8・4・26)バイトのテキストデータ（これをTX3'と呼ぶ）を作成し、テキストデータTX3'をN=4, M=26で分割する。本例では、その8文字分の部分テキストデータTG(i, j)を、次のようにテキストデータを単にアスキーコード（数値データ）に変換する関数asc()を用いて、そのまま64ビットの部分データAG(i, j)として扱う。

【0139】

$$AG(i, j) = asc(TG(i, j)) \quad \dots (43)$$

なお、図14にTG(3, 11)の変換例で示すように、関数asc(TG(i, j))は、部分テキストデータTG(i, j)の先頭の文字のコードが最下位桁となり、末尾の文字のコードが最上位桁となるように反転して変換を行う。なお、この際に、各アミノ酸を6ビットのデータで表してもよいが、データ量は3/4程度になるだけであるため、本例では部分テキストデータ（アスキーコード列）をそのまま部分データ（数値データ）として扱う。

【0140】

図15は、試料Gの部分データAG(i, j)の配列を示している。それに続いて図13の例と同様に、情報処理装置10は、その図15の4行で26列の64ビットの部分データAG(i, j)をガロア体GF(2<sup>64</sup>) (m=64)上の元のベクトル表示とみなして、部分データAG(i, j)に対してガロア体GF(2<sup>64</sup>)上の所定の演算を施す。ガロア体GF(2<sup>64</sup>)上の既約多項式GF(X

）及び生成元 $\alpha$ としては、(35)式(又は(35A)式)、及び(36)式(又は(36A)式)などを使用できる。

#### 【0141】

具体的に、情報処理装置10は、図15の各行( $i = 1 \sim 4$ )の部分データ $AG(i, j)$ に対してガロア体 $GF(2^{64})$ 上で、非配列方向( $j = 1 \sim 26$ )に対する和である第1パリティ(Parity) $B1G(i)$ 、 $\sum \alpha^{(j-1)} \cdot AG(i, j)$ である第2パリティ $B2G(i)$ 、及び $\sum \alpha^{2(j-1)} \cdot AG(i, j)$ である第3パリティ $B3G(i)$ を計算する。これらの計算式は、(19)式～(21)式に対応しており、この第1組のパリティ情報( $B1G(i) \sim B3G(i)$ )はそれぞれ64ビットである。

#### 【0142】

更に、情報処理装置10は、図15の各列( $j = 1 \sim 26$ )の部分データ $AG(i, j)$ に対してガロア体 $GF(2^{64})$ 上で、配列方向( $i = 1 \sim 4$ )に対する和である第1パリティ $C1G(j)$ 、 $\sum \alpha^{(i-1)} \cdot AG(i, j)$ である第2パリティ $C2G(j)$ 、及び $\sum \alpha^{2(i-1)} \cdot AG(i, j)$ である第3パリティ $C3G(j)$ を計算する。これらの計算式は、(22)式～(24)式に対応しており、この第2組のパリティ情報( $C1G(j) \sim C3G(j)$ )もそれぞれ64ビットである。このように計算して得られたパリティ $B1G(i) \sim B3G(i)$ 、 $C1G(j) \sim C3G(j)$ が、図15に16進数表示で示されている。

#### 【0143】

この例においては、各行、各列で3つのパリティを計算しているが、アミノ酸の配列はヌクレオチドの配列に比べるとデータ量がかなり少ないため、実用的には、パリティ情報は各行(非配列方向)、及び各列(配列方向)においてそれぞれ1つ(例えば $B2G(i)$ と $C2G(j)$ )としてもよい。図15の例において、パリティ $B2G(i)$ 、 $C2G(j)$ のみを使用するものとする、パリティはそれぞれ64ビット(8バイト)であるため、全部のパリティのデータ量は、240(=8・30)バイトとなる。従って、全部のパリティのデータ量は、全体の元のテキストデータTX3(820バイト)に対してほぼ1/3に減少している。

## 【 0 1 4 4 】

次に、情報処理装置 1 0 は、試料 G の名前の情報、配列の数 N A 3、テキストデータ T X 3、要約値 A B 3、A B R 3、及びパリティ情報を磁気ディスク装置 1 7 のワーキングファイル 2 0 に記録する。この際に、ワーキングファイル 2 0 を複数のファイルとしてもよい。その後、情報処理装置 1 0 は、試料 G の名前の情報、配列の数 N A 3、配列 S T 3、S B 3、要約値 A B 3、逆方向の配列 S T R 3、S B R 3、及び逆方向の要約値 A B R 3 を磁気ディスク装置 1 7 のコンテンツファイル 2 1 に記録する。更に、情報処理装置 1 0 は、コンテンツファイル 2 1 中の情報を通信ネットワーク 1 を介してコンテンツのプロバイダ 3 に送信する。これによって、コンテンツファイル 2 1 中の情報はプロバイダ 3 のサーバ内の閲覧可能なコンテンツファイル 3 1 に記録されて、第 3 者がインターネットを介して自由に閲覧できるようになる。この結果、第 3 者は、公開されている試料 G の配列の数 N A 3、及び要約値 A B 3（又は必要に応じて A B R 3）を自分の保有するアミノ酸の配列の配列数、及び要約値と比較することによって、その試料 G が自分にとって新規かどうかを判定できる。また、ユーザは、その試料 G の配列情報を複数の供給者から誤って重複して購入することを回避することができる。

## 【 0 1 4 5 】

その後、コンピュータシステム 2 A の所有者（アミノ酸情報の供給者）は、ユーザから購入要求が来るのを待つ状態となる。そして、ユーザから試料 G に対する簡易データの要求があったときには、情報処理装置 1 0 は、磁気ディスク装置 1 7 のワーキングファイル 2 0 の中の試料 G のパリティ情報（例えばその中の B 2 G ( i ) , C 2 G ( j ) ）を例えば電子メールの添付ファイルとしてそのユーザに送信する。パリティの情報を購入したユーザは、試料 G と同じ種類の自分で解読した試料のアミノ酸の配列のパリティと、その購入したパリティとを比較することによって、相違する部分の検出及び復元を或る程度行うことができる。

## 【 0 1 4 6 】

一方、ユーザから完全データの要求があったときには、情報処理装置 1 0 は、ワーキングファイル 2 0 中のテキストデータ T X 3 を Z I P ファイル等の形式で

圧縮し、この圧縮されたデータを例えば電子メールの添付ファイルとしてそのユーザに送信する。この際に必要に応じて、ハッシュ関数による要約値  $AB3$  を同時に送信してもよい。本例によれば、簡易データ（パリティ情報）はデータ量を少なくするために短時間で送信することができる。

## 【0147】

更に、そのアミノ酸の配列情報の供給者は、ワーキングファイル 20 に記録した情報、即ち試料  $G$  の名前の情報、配列の数  $NA3$ 、テキストデータ  $TX3$ 、要約値  $AB3$ 、 $ABR3$ 、及びパリティ情報を  $CD-R/RW$  ドライブ 15 を介して  $CD-R16$  に記録してもよい。この  $CD-R16$  から、更に多数の  $CD-ROM$  を作製してもよく、これらの記録媒体が郵送等によってユーザに販売される。

## 【0148】

なお、上記の実施の形態では、生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さが  $m$  ビット（ $m$  は 16 以上の整数）の部分データに分割し、複数行の前記部分データに各行毎に前記非配列方向にガロア体  $GF(2^m)$  上の第 1 の演算を施して第 1 組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体  $GF(2^m)$  上の第 2 の演算を施して第 2 組のパリティ情報を求めている。

## 【0149】

これに関して、 $m$  が 16 より小さい場合には、各部分データは、例えば 1～7 個程度のヌクレオチド、又は 1～2 個程度のアミノ酸の配列に対応するため、1 つの生物学的物質の配列に対して計算すべきパリティ情報の数が多くなり過ぎて好ましくない。また、最近のコンピュータの計算能力を活かしきれないという不都合もある。特に、コンピュータの処理単位が 64 ビットの倍数である場合には、パリティ情報の計算効率を高めるために、その  $m$  の値も 64, 128, 192, 256 などの 64 の倍数とすることが望ましい。

## 【0150】

また、 $m$ ビットを超える素数を  $p$  ( $> 2^m$ ) とすると、この素数  $p$  を法とするガロア体  $GF(p)$  を用いてパリティ情報を計算することも可能である。しかしながら、このガロア体  $GF(p)$  を用いた場合には、 $m$ ビットの部分データに演算を施して得られる個々のパリティ情報が  $m$ ビットを超える場合があるため、パリティ情報が必要以上に長くなるという不都合がある。これに対して、ガロア体  $GF(2^m)$  を用いた場合には、個々のパリティ情報を  $m$ ビットにできるため、パリティ情報を簡潔に記録できる利点がある。

## 【0151】

ここで、上記の実施の形態で使用するハッシュ関数に関して説明する。通常の暗号理論で使用されるハッシュ関数は、テキストデータ中のスペースコード及び改行コード等も全て演算処理対象としているが、ヌクレオチド及びアミノ酸の配列情報については見やすくするために、例えば配列番号1～3で示すように、途中にスペースコード、順序を示す数字コード、及び改行コードを挿入する場合がある。そこで、ヌクレオチドやアミノ酸などの生物学的物質の配列情報を演算処理対象とするハッシュ関数においては、テキストデータ中の所定コードとしてのスペースコード、数字コード、及び改行コードを無視する機能を付加することが望ましい。また、隣接する文字を”-”（ハイフン）で分けることも考えられるが、この場合には、更に”-”記号も無視する必要がある。また、例えばファイルの最後に「データの終わりを示すコード」が付加されるような場合には、そのコードも無視するようにしてもよい。

## 【0152】

また、例えばヌクレオチドの配列が、通常は小文字で表されるような場合には、ハッシュ関数に、選択的に大文字を小文字に変換して要約値を計算する機能を持たせるようにしてもよい。逆に、例えばアミノ酸の配列が、通常は大文字で表されるような場合には、ハッシュ関数に、選択的に小文字を大文字に変換して要約値を計算する機能を持たせるようにしてもよい。

## 【0153】

更に、原ファイルを複数の分割ファイルに分割する際には、複数の分割ファイルの順序等を示すデータ（以下、「コメントデータ」と言う）を各分割ファイル

に付加することが望ましいことがある。このように分割ファイル、又は1つの原ファイルにコメントデータを付加する場合にも、コメントデータはハッシュ関数で無視する必要がある。そのため、例えばコメントデータは所定の開始記号（例えば  $\backslash *$  ）及び終了記号（例えば  $* \backslash$  ）の間に記録し、ハッシュ関数で処理する際に開始記号から終了記号までのデータは無視するようにすればよい。

## 【 0 1 5 4 】

また、上記の実施の形態では、例えば生物のDNAのヌクレオチドの配列（又はタンパク質のアミノ酸の配列）内の先頭の一部、及び末尾の一部の配列、並びにその配列のテキストデータの要約値をインターネット上で公開することがある。この場合には、その公開されている一部の配列と、その要約値とからそのテキストデータの内容が推定される可能性もある。これを回避するために、そのテキストデータをハッシュ関数で処理する際に、その公開されている配列を除いた部分についてのみ、そのハッシュ関数を施して要約値を求めるようにしてもよい。

## 【 0 1 5 5 】

なお、本発明は上述の実施の形態に限定されず、本発明の要旨を逸脱しない範囲で種々の構成を取り得ることは勿論である。

## 【 0 1 5 6 】

## 【発明の効果】

本発明によれば、核酸や遺伝子中のヌクレオチド、又は又はタンパク質やペプチド中のアミノ酸などの生物学的物質の配列情報を、それらの配列を示すテキストデータよりも少ないデータ量のパリティ情報として近似的に記録することができる。従って、そのパリティ情報は、低容量の記録媒体にも記録できると共に、通信回線を介して短時間に送信することが可能となる。また、ガロア体GF(2<sup>m</sup>)上の演算を行うことによって、個々のパリティ情報を部分データと同じmビットの情報量で簡潔に記録できる利点がある。

## 【 0 1 5 7 】

また、2つの生物学的物質の配列のパリティ情報を比較することによって、2つの配列間の相違部の位置を少ないデータ量で容易に特定（検出）できると共に、必要に応じてその相違部の情報を復元することができる。従って、例えばSN

P（一塩基変位多型：Single Nucleotide Polymorphism）を少ないデータ量で容易に発見することができる。

【0158】

また、本発明によれば、ヌクレオチド又はアミノ酸などの生物学的物質の配列情報を近似する情報（パリティ情報）を少ないデータ量でユーザに供給できるビジネスモデルを提供することができる。この場合に、更に数学的な要約値を用いることによって、ユーザが提供された配列情報と情報供給者が保持している配列情報との同一性の確認などを容易に行うことができる。また、同一の複数の配列情報を誤って購入することも防止できる。

【図面の簡単な説明】

【図1】 本発明の実施の形態の一例で使用されるコンピュータシステムを示す概略構成図である。

【図2】 その実施の形態の一例で処理対象とするDNA、及びそのヌクレオチドの配列のバイナリーデータによる表現の例を示す図である。

【図3】 その実施の形態の一例におけるDNA情報の供給者の動作の一部を示すフローチャートである。

【図4】 図3の動作に続くDNA情報の供給者の動作を示すフローチャートである。

【図5】 その実施の形態の一例におけるDNA情報のユーザの動作の一部を示すフローチャートである。

【図6】 図5の動作に続くDNA情報のユーザの動作を示すフローチャートである。

【図7】 標準試料E（DNA）のヌクレオチド（2048個）の配列を表すテキストデータを4行で32列の部分テキストデータ $T(i, j)$ に分割した状態を示す図である。

【図8】 標準試料Eの部分データ $A(i, j)$ 、及びこれらから算出されるパリティ $B1(i) \sim C3(j)$ を示す図である。

【図9】 試料F（DNA）のヌクレオチド（2048個）の配列を表すテキストデータを4行で32列の部分テキストデータ $T_F(i, j)$ に分割した状

態を示す図である。

【図10】 試料Fの部分データ  $AF(i, j)$ 、及びこれらから算出されるパリティ  $B1F(i) \sim C3F(j)$  を示す図である。

【図11】 標準試料Eのパリティと異なる試料Fのパリティ、及び復元された部分データを示す図である。

【図12】 未知数  $X1, X2, Y1, Y2$  をガロア体  $GF(2^{128})$  上で求める場合の計算を示すフローチャートである。

【図13】 図7の標準試料Eのヌクレオチドの配列を表すテキストデータをバイナリーデータに変換した後、5行で13列の部分データ  $B(i, j)$  に分割した状態を示す図である。

【図14】 試料G（タンパク質）のアミノ酸（820個）の配列を表すテキストデータを4行で26列の部分テキストデータに分割した状態を示す図である。

【図15】 図14のアミノ酸の配列に対して計算されたパリティ  $B1G(i) \sim C3G(j)$  を示す図である。

#### 【符号の説明】

1…通信ネットワーク、2A, 2B…コンピュータシステム、3…コンテンツのプロバイダ、4…DNAのシーケンサー、10…情報処理装置、15…CD-R/RWドライブ、16…CD-R、17…磁気ディスク装置、19…マスターファイル、20…ワーキングファイル、21…コンテンツファイル、31…コンテンツファイル

#### 【配列表】

SEQUENCE LISTING

<110> Omori, Satoshi



<120> Method and apparatus for recording sequence information  
of biological materials

<130> 2001A16

<140>

<141>

<150> PCT/JP01/03324

<151> 2001-04-18

<160> 3

<170> PatentIn Ver. 2.0

<210> 1

<211> 2048

<212> DNA

<213> Escherichia coli

<400> 1

agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc 60

tgatagcagc ttctgaactg gttacctgcc gtgagtaaata taaaatttta ttgacttagg 120  
 tcactaaata ctttaaccaa tataggcata gcgcacagac agataaaaat tacagagtac 180  
 acaacatcca tgaaacgcat tagcaccacc attaccacca ccatcaccat taccacaggt 240  
 aacggtgcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg 300  
 cttttttttt cgaccaaagg taacgaggta acaaccatgc gagtgttgaa gttcggcggt 360  
 acatcagtgg caaatgcaga acgttttctg cgtgttgccg atattctgga aagcaatgcc 420  
 aggcaggggc aggtggccac cgtcctctct gccccgccca aaatcaccaa ccacctggtg 480  
 gcgatgattg aaaaaacat tagcggccag gatgctttac ccaatatcag cgatgccgaa 540  
 cgtatttttg ccgaactttt gacgggactc gccgccgcc agccgggggtt cccgctggcg 600  
 caattgaaaa ctttcgtcga tcaggaattt gcccaaataa aacatgtcct gcatggcatt 660  
 agtttggttg ggcagtgcgc ggatagcatc aacgctgcgc tgatttgccg tggcgagaaa 720  
 atgtcgatcg ccattatggc cggcgtatta gaagcgcgcg gtcacaacgt tactgttata 780  
 gatccggtcg aaaaactgct ggcagtgggg cattacctcg aatctaccgt cgatattgct 840  
 gagtccaccc gccgtattgc ggcaagccgc attccggctg atcacatggt gctgatggca 900  
 ggtttcaccg ccggtaatga aaaaggcgaa ctggtggtgc ttggacgcaa cggttccgac 960

tactctgctg cggctgctggc tgcctgttta cgcgccgatt gttgagagat ttggacggac 1020

gttgacgggg tctataacctg cgacccgcgt cagggtcccc atgcgaggtt gttgaagtcg 1080

atgtcctacc aggaagcgat ggagctttcc tacttcggcg ctaaagttct tcacccccgc 1140

accattaccc ccatcgccca gttccagatc cttgcctga ttaaaaatac cggaaatcct 1200

caagcaccag gtacgctcat tgggtgccagc cgtgatgaag acgaattacc ggtcaagggc 1260

attccaatc tgaataacat ggcaatgttc agcgtttctg gtccggggat gaaagggatg 1320

gtcggcatgg cggcgcgcggt ctttgcagcg atgtcacgcg cccgtatttc cgtggtgctg 1380

attacgaat catcttccga atacagcatc agtttctgcg ttccacaaag cgactgtgtg 1440

cgagctgaac gggcaatgca ggaagagttc tacctggaac tgaaagaagg ctactggag 1500

ccgctggcag tgacggaacg gctggccatt atctcggtgg taggtgatgg tatgcgacc 1560

ttgcgtggga tctcggcgaa attctttgcc gcaactggccc gcgccaatat caacattgtc 1620

gccattgctc agggatcttc tgaacgctca atctctgtcg tggtaaataa cgatgatgcg 1680

accactggcg tgcgcgttac tcatcagatg ctgttcaata ccgatcaggt tatcgaagtg 1740

tttgtgattg gcgtcggtgg cgttggcggg gcgctgctgg agcaactgaa gcgtcagcaa 1800

agctggctga agaataaaca tatcgactta cgtgtctgcg gtgttgccaa ctcgaaggct 1860

ctgctcacca atgtacatgg ccttaatctg gaaaactggc aggaagaact ggcgcaagcc 1920

aaagagccgt ttaatctcgg gcgcttaatt cgcctcgtga aagaatatca tctgctgaac 1980

ccggtcattg ttgactgcac ttccagccag gcagtggcgg atcaatatgc cgacttcctg 2040

cgcggaagg 2048

<210> 2

<211> 2048

<212> DNA

<213> Escherichia coli

<400> 2

agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc 60

tgatagcagc ttctgaactg gttacctgcc gtgagtaaata taaaatttta ttgacttagg 120

tcactaaata cttaaccaa tataggcata gcgcacagac agataaaaat tacagagtac 180

acaacatcca tgaaacgcat tagcaccacc attaccacca ccatcaccat taccacaggt 240

aacggtgcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg 300

cttttttttt cgaccaaagg taacgaggta acaaccatgc gagtggtgaa gttcggcgggt 360  
 acatcagtgg caaatgcaga acgttttctg cgtgttgccg atattctgga aagcaatgcc 420  
 aggcaggggc aggtggccac cgtcctctct gccccgccca aaatcaccaa ccacctgggtg 480  
 gcgatgattg aaaaaacat tagcggccag gatgctttac ccaatatcag cgatgccgaa 540  
 cgtatttttg ccgaactttt gacgggactc gccgccgcc agccgggggtt cccgctggcg 600  
 caattgaaaa ctttcgtcga tcaggaattt gcccaaataa aacatgtcct gcatggcatt 660  
 agtttggttg ggcagtgcc gccatagcatc aacgctgcgc tgatttgccg tggcgagaaa 720  
 atgtcgatcg ccattatggc cggcgtatta gaagcgcgcg gtcacaacgt tactgttacc 780  
 gatccggtcg aaaaactgct ggcagtgggg cattacctcg aatctaccgt cgatattgct 840  
 gagtccaccc gccgtattgc ggcaagccgc attccggctg atcacatggt gctgatggca 900  
 ggtttcaccg ccggtaatga aaaaggcgaa ctggtgggtg ttggacgcaa cggttccgac 960  
 tactctgctg cggtgctggc tgcctgttta cgcgccgatt gttgcgagat ttggacatta 1020  
 tggcggccaa cttataacct cgacccgcgt cagggtgccg atgcgagggt gttgaagtcg 1080  
 atgtcctacc aggaagcgat ggagctttcc tacttcggcg ctaaagtctc tcacccccgc 1140

accattaccc ccatcgccca gttccagatc ccttgccctga ttaaaaatac cggaatcct 1200

caagcaccag gtacgtcat tgggtgccagc cgtgatgaag acgaattacc ggtcaagggc 1260

atttccaatc tgaataacat ggcaatgttc agcgtttctg gtccggggat gaaagggatg 1320

gtcggcatgg cggcgcgcg ttttcagcgc atgtcacgcg cccgtatttc cgtggtgctg 1380

attacgcaat catcttccga atacagcatc agtttctgcg ttccacaaag cgactgtgtg 1440

cgagctgaac gggcaatgca ggaagagttc tacctggaac tgaaagaagg cttactggag 1500

ccgctggcag tgacggaacg gctggccatt atctcggtgg taggtgatgg tatgcgacc 1560

ttgcgtggga tctcggcgaa attctttgcc gcactggccc gcgccaatat caacattgtc 1620

gccattgctc aggatcttc tgaacgctca atctctgtcg tggtaaataa cgatgatgcg 1680

accactggcg tgcgcgttac tcatcagatg ctgttcaata ccgatcaggt tatcgaagtg 1740

tttgtgattg gcgtcggtgg cgttggcggt gcgctgctgg agcaactgaa gcgtcagcaa 1800

agctggctga agaataaaca tatcgactta cgtgtctgcg gtgttgccaa ctcgaaggct 1860

ctgctcacca atgtacatgg ccttaatctg gaaaactggc aggaagaact ggcgcaagcc 1920

aaagagccgt ttaatctcgg gcgcttaatt cgcctcgtga aagaatatca tctgctgaac 1980

ccggtcattg ttgactgcac ttccagccag gcagtggcgg atcaatatgc cgacttcctg 2040

cgCgaagg

2048

<210> 3

<211> 820

<212> PRT

<213> Escherichia coli

<400> 3

Met Arg Val Leu Lys Phe Gly Gly Thr Ser Val Ala Asn Ala Glu Arg

1

5

10

15

Phe Leu Arg Val Ala Asp Ile Leu Glu Ser Asn Ala Arg Gln Gly Gln

20

25

30

Val Ala Thr Val Leu Ser Ala Pro Ala Lys Ile Thr Asn His Leu Val

35

40

45

Ala Met Ile Glu Lys Thr Ile Ser Gly Gln Asp Ala Leu Pro Asn Ile

50

55

60

Ser Asp Ala Glu Arg Ile Phe Ala Glu Leu Leu Thr Gly Leu Ala Ala

65

70

75

80

Ala Gln Pro Gly Phe Pro Leu Ala Gln Leu Lys Thr Phe Val Asp Gln

85

90

95

Glu Phe Ala Gln Ile Lys His Val Leu His Gly Ile Ser Leu Leu Gly

100

105

110

Gln Cys Pro Asp Ser Ile Asn Ala Ala Leu Ile Cys Arg Gly Glu Lys

115

120

125

Met Ser Ile Ala Ile Met Ala Gly Val Leu Glu Ala Arg Gly His Asn

130

135

140

Val Thr Val Ile Asp Pro Val Glu Lys Leu Leu Ala Val Gly His Tyr

145

150

155

160

Leu Glu Ser Thr Val Asp Ile Ala Glu Ser Thr Arg Arg Ile Ala Ala

165

170

175

Ser Arg Ile Pro Ala Asp His Met Val Leu Met Ala Gly Phe Thr Ala



180

185

190

Gly Asn Glu Lys Gly Glu Leu Val Val Leu Gly Arg Asn Gly Ser Asp

195

200

205

Tyr Ser Ala Ala Val Leu Ala Ala Cys Leu Arg Ala Asp Cys Cys Glu

210

215

220

Ile Trp Thr Asp Val Asp Gly Val Tyr Thr Cys Asp Pro Arg Gln Val

225

230

235

240

Pro Asp Ala Arg Leu Leu Lys Ser Met Ser Tyr Gln Glu Ala Met Glu

245

250

255

Leu Ser Tyr Phe Gly Ala Lys Val Leu His Pro Arg Thr Ile Thr Pro

260

265

270

Ile Ala Gln Phe Gln Ile Pro Cys Leu Ile Lys Asn Thr Gly Asn Pro

275

280

285

Gln Ala Pro Gly Thr Leu Ile Gly Ala Ser Arg Asp Glu Asp Glu Leu

290

295

300

Pro Val Lys Gly Ile Ser Asn Leu Asn Asn Met Ala Met Phe Ser Val

305

310

315

320

Ser Gly Pro Gly Met Lys Gly Met Val Gly Met Ala Ala Arg Val Phe

325

330

335

Ala Ala Met Ser Arg Ala Arg Ile Ser Val Val Leu Ile Thr Gln Ser

340

345

350

Ser Ser Glu Tyr Ser Ile Ser Phe Cys Val Pro Gln Ser Asp Cys Val

355

360

365

Arg Ala Glu Arg Ala Met Gln Glu Glu Phe Tyr Leu Glu Leu Lys Glu

370

375

380

Gly Leu Leu Glu Pro Leu Ala Val Thr Glu Arg Leu Ala Ile Ile Ser

385

390

395

400

Val Val Gly Asp Gly Met Arg Thr Leu Arg Gly Ile Ser Ala Lys Phe

405

410

415

Phe Ala Ala Leu Ala Arg Ala Asn Ile Asn Ile Val Ala Ile Ala Gln

420

425

430

Gly Ser Ser Glu Arg Ser Ile Ser Val Val Val Asn Asn Asp Asp Ala

435

440

445

Thr Thr Gly Val Arg Val Thr His Gln Met Leu Phe Asn Thr Asp Gln

450

455

460

Val Ile Glu Val Phe Val Ile Gly Val Gly Gly Val Gly Gly Ala Leu

465

470

475

480

Leu Glu Gln Leu Lys Arg Gln Gln Ser Trp Leu Lys Asn Lys His Ile

485

490

495

Asp Leu Arg Val Cys Gly Val Ala Asn Ser Lys Ala Leu Leu Thr Asn

500

505

510

Val His Gly Leu Asn Leu Glu Asn Trp Gln Glu Glu Leu Ala Gln Ala

515

520

525

Lys Glu Pro Phe Asn Leu Gly Arg Leu Ile Arg Leu Val Lys Glu Tyr

530

535

540

His Leu Leu Asn Pro Val Ile Val Asp Cys Thr Ser Ser Gln Ala Val

545

550

555

560

Ala Asp Gln Tyr Ala Asp Phe Leu Arg Glu Gly Phe His Val Val Thr

565

570

575

Pro Asn Lys Lys Ala Asn Thr Ser Ser Met Asp Tyr Tyr His Gln Leu

580

585

590

Arg Tyr Ala Ala Glu Lys Ser Arg Arg Lys Phe Leu Tyr Asp Thr Asn

595

600

605

Val Gly Ala Gly Leu Pro Val Ile Glu Asn Leu Gln Asn Leu Leu Asn

610

615

620

Ala Gly Asp Glu Leu Met Lys Phe Ser Gly Ile Leu Ser Gly Ser Leu

625

630

635

640

Ser Tyr Ile Phe Gly Lys Leu Asp Glu Gly Met Ser Phe Ser Glu Ala

645

650

655

Thr Thr Leu Ala Arg Glu Met Gly Tyr Thr Glu Pro Asp Pro Arg Asp

660

665

670

Asp Leu Ser Gly Met Asp Val Ala Arg Lys Leu Leu Ile Leu Ala Arg

675

680

685

Glu Thr Gly Arg Glu Leu Glu Leu Ala Asp Ile Glu Ile Glu Pro Val

690

695

700

Leu Pro Ala Glu Phe Asn Ala Glu Gly Asp Val Ala Ala Phe Met Ala

705

710

715

720

Asn Leu Ser Gln Leu Asp Asp Leu Phe Ala Ala Arg Val Ala Lys Ala

725

730

735

Arg Asp Glu Gly Lys Val Leu Arg Tyr Val Gly Asn Ile Asp Glu Asp

740

745

750

Gly Val Cys Arg Val Lys Ile Ala Glu Val Asp Gly Asn Asp Pro Leu

755

760

765

Phe Lys Val Lys Asn Gly Glu Asn Ala Leu Ala Phe Tyr Ser His Tyr

770

775

780

Tyr Gln Pro Leu Pro Leu Val Leu Arg Gly Tyr Gly Ala Gly Asn Asp

785

790

795

800

Val Thr Ala Ala Gly Val Phe Ala Asp Leu Leu Arg Thr Leu Ser Trp

805

810

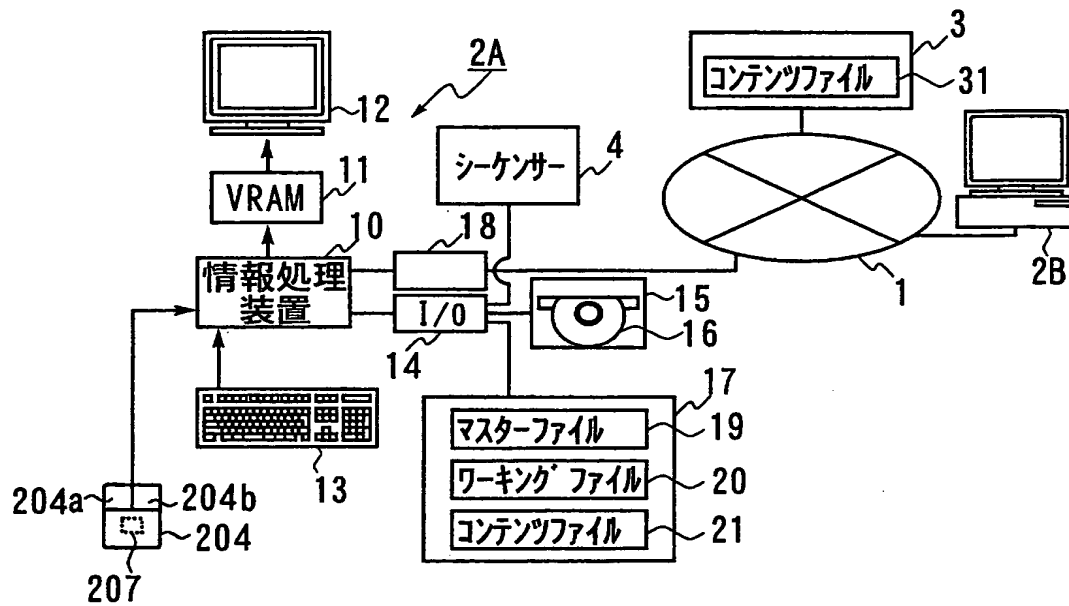
815

Lys Leu Gly Val

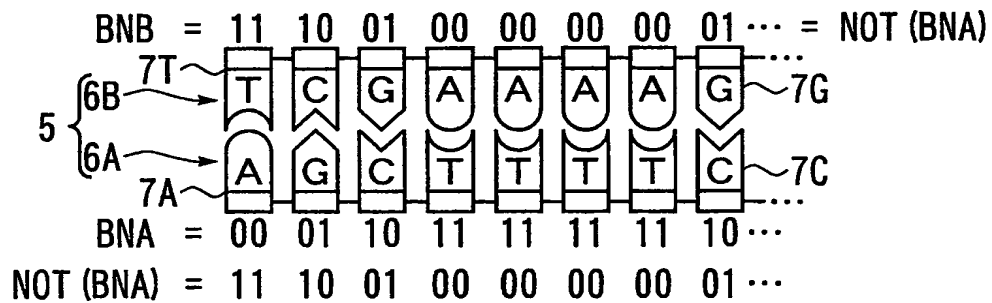
820

【書類名】 図面

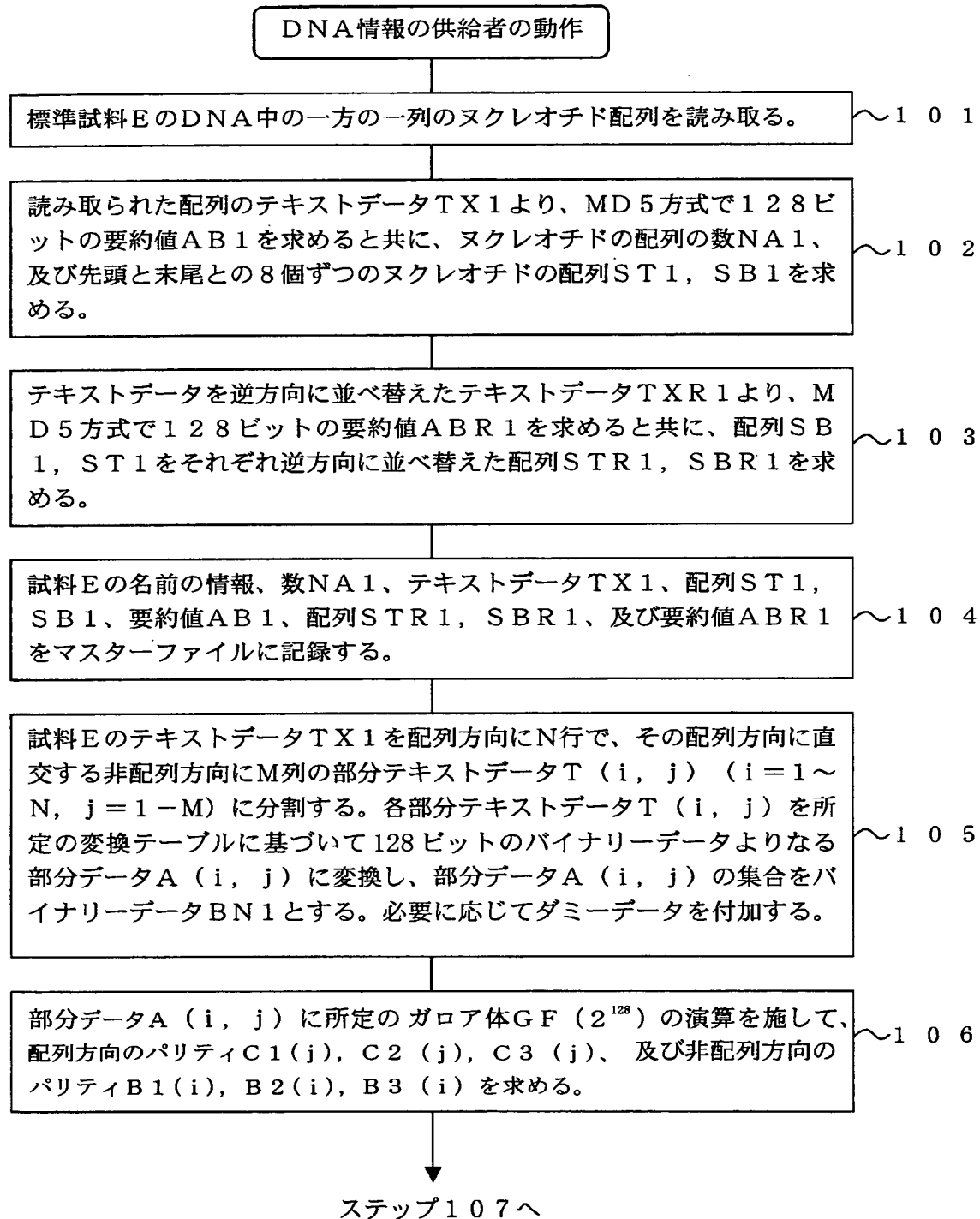
【図 1】



【図 2】

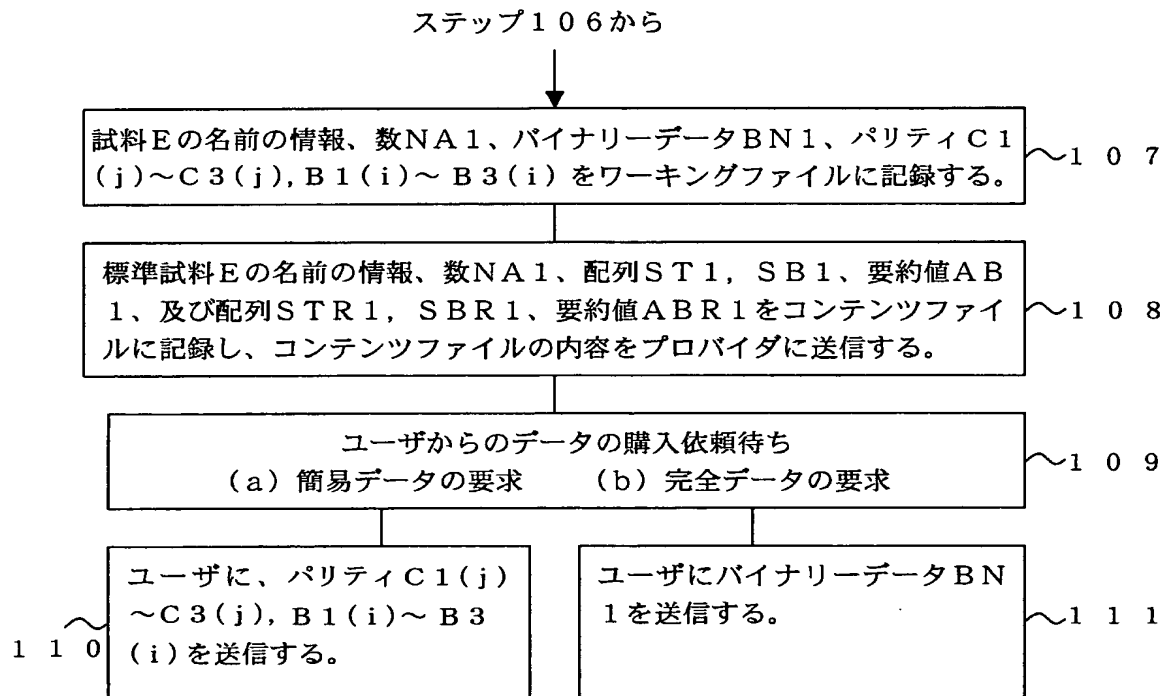


【図 3】

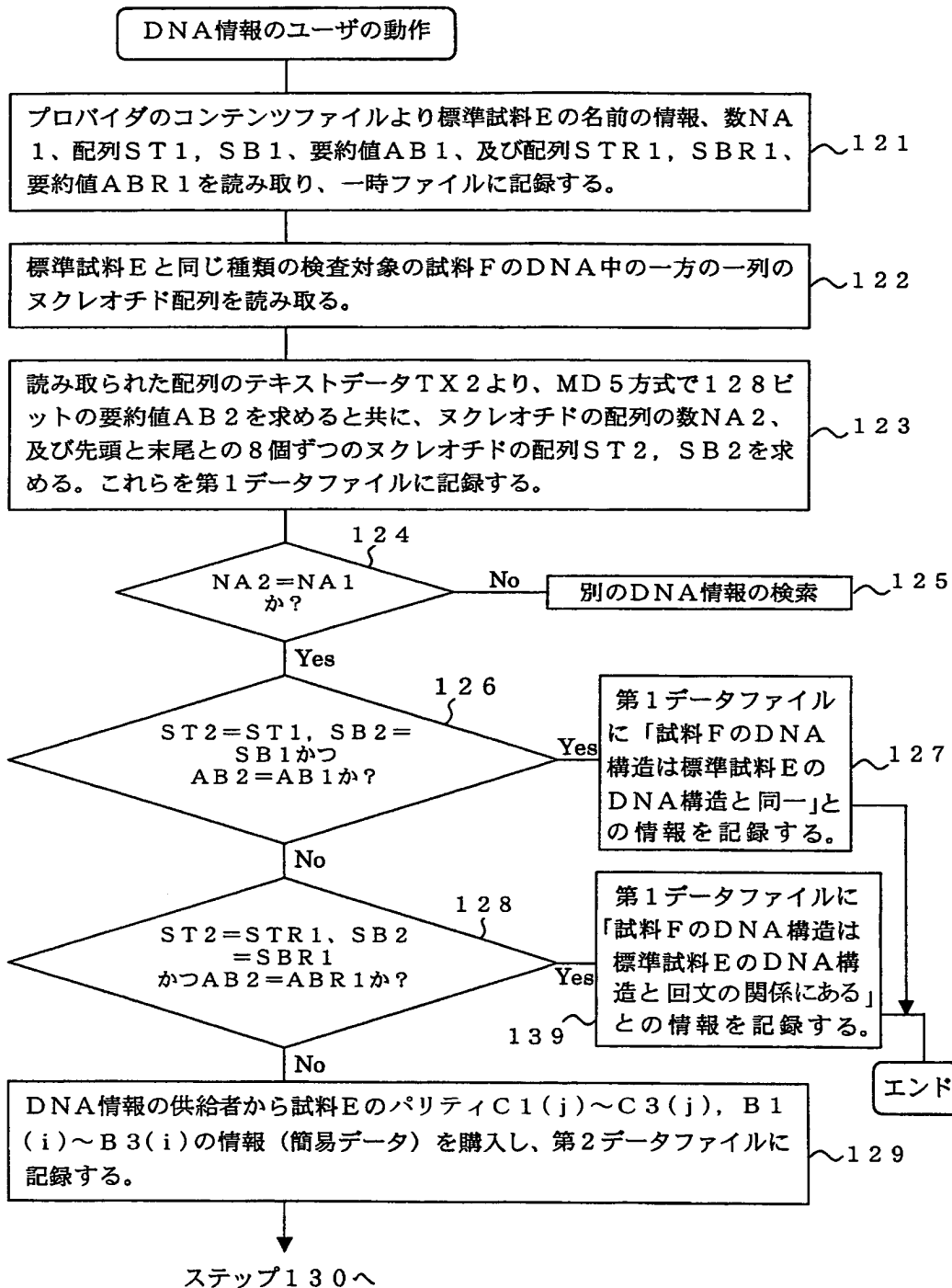




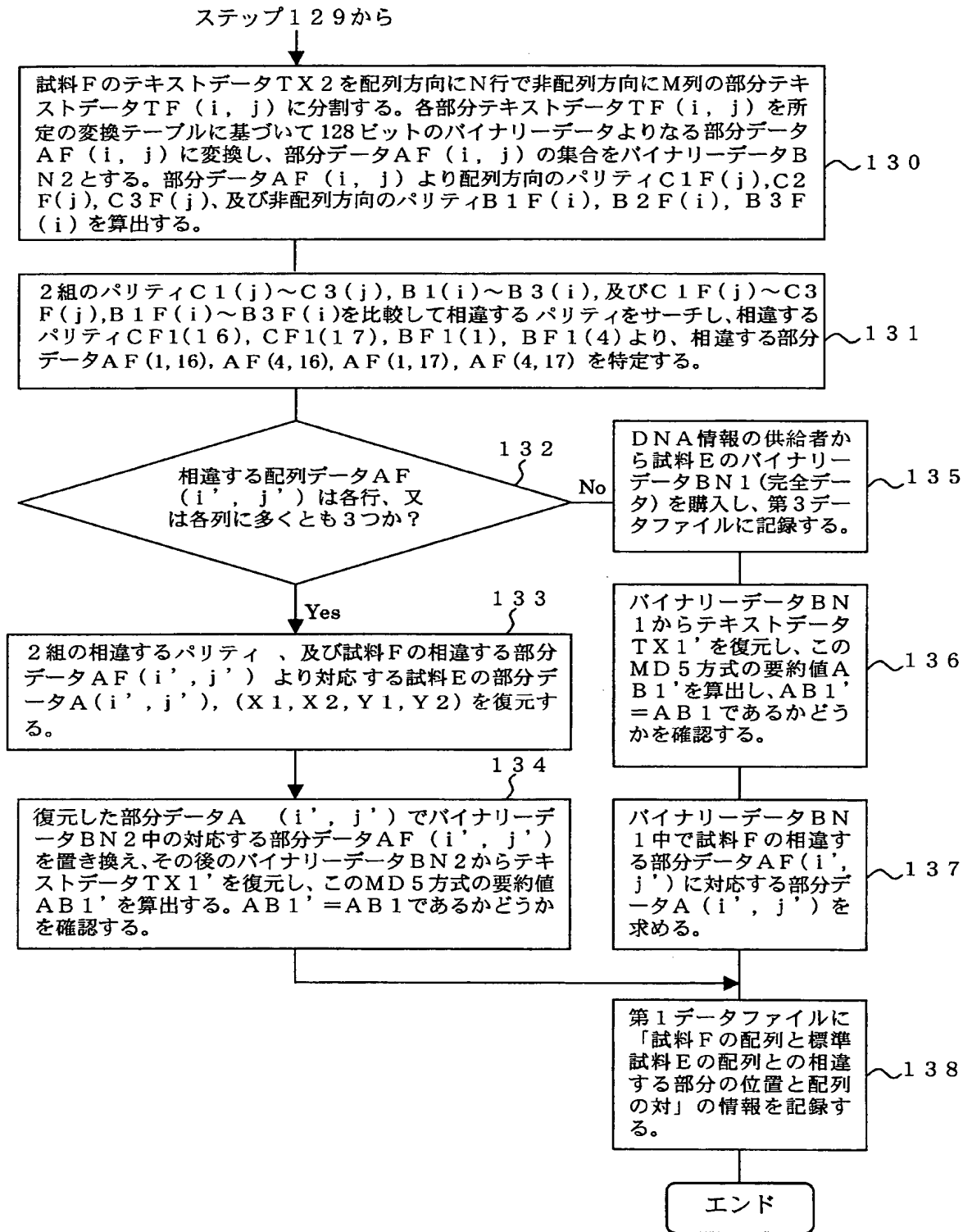
【図 4】



【図 5】



【図 6】



【図7】

標準試料E				$T(i, j)$	
$i$	$1$	$2$	$3$	$4$	
1	AGCTTTTCATTCTGAC	TGCAACGGGCAATATG	TCTCTGTGTGGATTAA	AAAAAGAGTGTCTGAT	
2	AGCAGCTTCTGAACGTG	GTTACCTGCGGTGAGT	AAATTAAATTTTATT	GACTTAGGTCACTAAA	
3	TACTTTAAACCAATATA	GGCATAGCGCACAGAC	AGATAAAATTTACAGA	GTACACAACATCCATG	
4	AAACGCATTAGCACCA	CCATTACCAACCACCAT	CACCATTTACACAGGT	AACGGTGGGGCTGAC	
5	GGTACAGGAACACACA	GAAAAAGCCCGCAC	TGACAGTGGGGCTTT	TTTTTTCGACCAGAGG	
6	TAACGAGGTAACAACC	ATCGAGTGTGAAAGT	TGCGCGGTACATCAGT	GGCAATGCAGAACGT	
7	TTTCTGGTGTGGCG	ATATTCTGAAAGCAA	TGCCAGGACGGGCAG	GTGGCCACCGTCTCT	
8	CTGCCCCGGCCAAAT	CACCAACCACTGGTG	GGCATGATTGAAAAA	CCATTAGCGGCCAGGA	
9	TGCTTTACCCAATATC	AGCGATGCCGAACGTA	TTTTTGGCAACTTTT	GACGGGACTCGCCGCC	
10	GCCAGCCGGGGTTCC	CGCTGGCGCAATTGAA	AACTTCGTGATCAG	GAATTTGCCCAATAA	
11	AACATGCTCTGCATGG	CATTAGTTTGTGGGG	<b>CAGTCCCCGATAGCA</b>	TCAACGCTGGCTGAT	
12	TTGCCGTGGCGAGAAA	ATGTCGATCGCCATTA	TGCCCGCGGTATTAGA	AGCGCGCGTCAACAC	
13	GTTACTGTTATCGATC	CGGTCGAAAAACTGCT	GGCAGTGGGCATTAC	CTCGAATCTACCGTCG	
14	ATATTGCTGAGTCCAC	CCGCCGATTGCGGCA	AGCCGCATTTCCGGTG	ATCACATGGTGTGAT	
15	GGCAGGTTTCACCGCC	GGTAATGAAAAAGCG	AACTGGTGGTGTGG	ACGCAACGGTTCGAC	
16	TACTCTGCTGCGGTGC	TGGCTGCTGTTTACG	CGCCGATTGTTGGAG	ATTTGGACGGACGTTG	
17	ACGGGGTCTATACCTG	CGACCCCGGTCAGGTG	CCCGATGCGAGGTTGT	TGAAGTCGATGTCCTA	
18	CCAGGAAGCGATGGAG	CTTTCCTACTTCGGCG	CTAAAGTTCTTCAACC	CCGCACCATTACCCCC	
19	ATCGCCAGTTCAGAG	TCCCTTGCTTGATTAA	AAATACCGGAAATCCT	CAAGCACCAAGGTACGC	
20	TCATTGGTGCCAGCCG	TGATGAAGACGAATTA	CCGGTCAAGGCAATT	CCAATCTGAATAACAT	
21	GGCAATGTTACAGCTT	TCTGGTCCGGGATGA	AAGGGATGGTCGGCAT	GGCGGCGCGGCTCTT	
22	GCAGCGATGTCACGCG	CCCGTATTTCGTGGT	GCTGATTACGCAATCA	TCTTCCGAATACAGCA	
23	TCAGTTCTGCGTTCC	ACAAAGCGACTGTGTG	CGAGCTGAACGGGCAA	TGCAGGAAGAGTTCTA	
24	CCTGGAACCTGAAAGAA	GGCTTACTGGAGCCGC	TGGCAGTGACGGAAACG	GCTGGCCATTATCTCG	
25	GTGGTAGGTGATGGTA	TGCGCACCTTGGCTGG	GATCTCGCGGAAATTC	TTTGGCGCACTGGCCC	
26	GCGCCAATATCAACAT	TGTCGCCATTGCTCAG	GGATCTTCTGAACGCT	CAATCTCTGTCTGGT	
27	AAATAACGATGATGCG	ACCACTGGCGTGGCGG	TTACTCATCAGATGCT	GTTCAATACCGATCAG	
28	GTTATCGAAGTGTG	TGATTGGCGTGGGTGG	CGTTGGCGGTGGCTG	CTGGAGCAACTGAAGC	
29	GTCAGCAAGCTGGCT	GAAGAATAACATATC	GACTTACGTGTCTGCG	GTGTTGCCAATCTGAA	
30	GGCTCTGCTCACC AAT	GTACATGGCTTAATC	TGGAATAACTGGCAGGA	AGAACTGGCGCAAGCC	
31	AAAGAGCCGTTAATC	TGGGGCGCTTAATTCG	CCTCGTGAAGAATAT	CATCTGCTGAACCCGG	
32	TCATTGTTGACIGCAC	TTCCAGCCAGGCAGTG	GCGGATCAATATGCCG	ACTTCTTGGCGGAAGG	

$A(3, 11) = \text{asc}(T(3, 11))$   
 $= \text{asc}(CAGTCCCCGATAGCA) = \text{hex}(414347415441474743434754474143)$

【図8】

標準試験E						
j	i	2	3	4		
A(i, j)						
1	1	1115150000617060406170017150200	ee6ae0fde69f53cea81f58dd19eeeb3	C1(j)	C2(j)	C3(j)
2	1	06060206021500151506000400151300	3081ce3ca2929962e5c3ca3230f4e0f1	2c54d2518734f909f5cfc45345a3f0a	195d06098541c84b8cc10e1a10d5390	77518b58c115a9a08189540c50f3985
3	1	04060615000015060206170002001515	7c9c0c1d8309e2dc8c0c9f39dca70e4	7e3ef48a38fa9b59ddc13d3f6a00820	0db502a78e8e82fffd5274570432607	f3fa80214bd4748d736e6036eaf9f4ca
4	1	02040017020206111511021513000200	28ae98c2d8c0c0c6ee887e6479627c3b	9ceadcc496d6ec58476d4cedc8d6bc	50876afe034e68be36e0cf7b061cb	34d59816d48bead7db8a66174a6b5e3
5	1	1113150200060606001711502131100	51d6f0c0f6daeae8efad8326d6c1c3	0eb7a499d41cd27ad2c8dbdb18e92bc	2991c9cda1809f2aaa39alc5dcd4ecd	5a076c115266a7dcbac0a0c462a1b87
6	1	17040202111317110013060206021106	52ce75a7e0d28acaf233ac18c2c2f1	eaecfd68a0d0dccc44ebf1b21cd69	283fe85987d2ecb336e0bef404368566	b83455d7588a0be1f214feabf6f1a04c
7	1	15001704110606110011000213111306	58fee68e9cde3ee6d8d1e0d2dcedb8c	58cd3c84dac1d489d17a81d2d6cd65c7	0c9908828ab56535f0275d6bd4a82de6	5559bf3afcbb3482c0163ef4d3db3d06
8	1	13130006170200151706040215021504	fa9cfa72d5fe4684bd4cccd5d5d86dc9	4ed2a286d67c321f19f18d288dcd8c7	054b34f2b306cfa0db5f8c3787f3869	b160cbc121f18146caae91f86e27bf60
9	1	15171500000606130004000600171506	6e66948e80a2db8e9ae5cc7d6270fa	164386867c154ecd12d32fed6c2f4b9	86a8be9ebc00476b5784e80ed3d586a3	2ba6f503438cc16e8ba76d6de0a1ddc
10	1	04020415130006130004000617020402	5ad0ded6c5b21c6a9e2c0cef0d6a0d3	0cf2feb066e5c3d6c0daecf4aacaddf7	a3078606253a6a67884a21d2160baee	b4832e73fb9cbefddedc0f9f8b78fdd
11	1	15021313001517170000041100110215	b2b0a0cac472d2e25e2d229ebd4daea	64daa2e4d0c26d8f3828e147634d27f	6abace598213d72418d6d33a18af27	2716a8643fb2f951a7ddcb56681ef7c
12	1	02131511170604041711000013040000	a17d40c676f886fdd5aae3e8e8ecfd1	464ccd4376ccc6e9c2c5b1e1d23ce278	12b4e8d38657b4d37d8e1d3f73fac2	81dc839f9858b261efeealcba1b487b
13	1	13150600021506061115060613130000	64daa2e4d0c26d8f3828e147634d27f	64d866d870d4667dfc8cc4d086dcaa7	a8atef35889f618174dc18be1a527abd	f62afa0d19b28da94b5a7558b04c18d9
14	1	11170017130402000602021315060402	5c91a4f6ed88e8fbc6e6d499fe327d51	4ab1fbcdf6da4d8d0deaeab82bd65f9	c6dd616cbb3c827868bffa2184c02dfb	4577e7874d6750e0f5861209a754f7
15	1	00061313021302151504150017130200	adfae246f8424382798c34da36e0dae9	41ece2a4798ed298c4fefe363c1671ef	a5b5e00276b5ebccbb9b01821d683b347	f39c84d4957540d74afadb38ba0828fd
16	1	0411061717021300171151700131502	26c2ce21c0c8c5092baee9cc18034a9	49eee4ceecceee299faf65bdeedd6c893	01f43a66cd9ac887987cce903828f629	
17	1	151313131700150417040202040015	4ab1fbcdf6da4d8d0deaeab82bd65f9	22eda2f6c9c1f34062caaa70de3cc0ff		
18	1	00020002170013171302060411130000	64d866d870d4667dfc8cc4d086dcaa7	6edeca2b9d2a4daedf7d82c96d78fcea		
19	1	17021502171506020604151517001717	5c91a4f6ed88e8fbc6e6d499fe327d51			
20	1	13041317040204110006111513020017	64daa2e4d0c26d8f3828e147634d27f			
21	1	06020006021706001513061306060400	adfae246f8424382798c34da36e0dae9			
22	1	150604061302131317020606130215	26c2ce21c0c8c5092baee9cc18034a9			
23	1	13041317040204110006111513020017	4ab1fbcdf6da4d8d0deaeab82bd65f9			
24	1	040213131706130411001100020002	41ece2a4798ed298c4fefe363c1671ef			
25	1	02061100150617061115041517040017	49eee4ceecceee299faf65bdeedd6c893			
26	1	13040002040613060215020400130217	22eda2f6c9c1f34062caaa70de3cc0ff			
27	1	130204170613110215111715170213	6edeca2b9d2a4daedf7d82c96d78fcea			
28	1	040002150404000400040606060013				
29	1	11150604021700150415040606060000				
30	1	15110002151100000006150017041315				
31	1	17110217171300151500171500060015				
32	1	04110600040615041100171704111706				
B1(i)	1	111113131306021700150200151517	041104171513151310215040402150204			
B2(i)	1	0402041502001511	1313040002021304	0402131504041511	1100041317040000	
B3(i)	1	b9552e59d8b8f1bf	59d9c9ced8917afd	7b9c721c2814474	8254675fde5e039c	
	2	cb829a50cfeaac8e	825d71630ac6470	1e3fc85c3787ec0e	e2f606c26e522789	
	3	ce9dc74c4686b8f2	03959294b8c2898f	1fhe211ff9a71f23	ac32e1ddc4db0e9b	
	4	86cdfa380f1e17b8	c659b84602915791	54f20aa364036e0c	032628431bb00604	

【図9】

試料 F					TF (i, j)	
i	1	2	3	4		
j	1	2	3	4		
1	AGCTTTTCATTCTGAC	TGCAACCGGGCAATATG	TCTCTGTGTGGATTAA	AAAAAGAGTGTCTGAT		
2	AGCAGCTTCTGAACTG	GTTACCTGCGGTGAGT	AAATTAAATTTTATT	GACTTAGGTCACATAA		
3	TACTTTAACCAATATA	GGCATAGCGCACAGAC	AGATAAAATTTACAGA	GTACACAACATCCATG		
4	AAACGCATTAGCACCA	CCATTACCAACCACAT	CACCATTACCACAGGT	AACGGTCCGGGCTGAC		
5	GCGTACAGGAACACA	GAAAAAGCCCGCAC	TGACAGTCCGGCTTT	TTTTTTCGACCAAGG		
6	TAACGAGGTAACAACC	ATGCGAGTGTGAAGT	TGCGCGGTACATCAGT	GGCAAAATGCAGAACGT		
7	TTTCTGCGTGTGCGG	ATATTCTGGAAGCAA	TGCCAGGCAAGGGCAG	GTGGCCACCGTCTCT		
8	CTGCCCGGCCAAAT	CACCAACCACCTGGTG	GCGATGATTGAAAAA	CCATTAGCGGCCAGGA		
9	TGCTTTACCCCAATATC	AGCGATGCCGAACGTA	TTTTTCCGAACITTT	GACGGGACTCGCCGCC		
10	GCCACGCCGGGTTC	CGCTGGCGCAATTGAA	AACTTCTGTCGATCAG	GAATTTGCCCAATAA		
11	AACATGCTCTGCATGG	CATTAGTTTGTGGGG	<u>CAGTCCCGGATAGCA</u>	TCAACGCTGCGCTGAT		
12	TTGCCGTGGCGAGAAA	ATGTCGATCGCCATTAA	TGGCCGGCGTATTAGA	AGCGCGGGTCACAAC		
13	GTTACTGTTATCGATC	CGGTCGAAAAACTGCT	GGCAGTGGGGCATTAC	CTCGAATCTACCGTCG		
14	ATATTGCTGAGTCCAC	CCGCCGTATTGCGGCA	AGCCGCATTCGGGCTG	ATCACATGGTGTGAT		
15	GGCAGGTTTCACCGCC	GGTAATGAAAAAGCGG	AACCTGGTGGTCTTGG	ACGCAACGGTTCGAC		
16	TACTCTGCTGCGGTGC	TGGCTGCCCTGTTTACG	CGCCGATTGTTGCGAG	ATTTGGACATTATGGC		
17	GGCCAACTTATACCTG	CGACCCCGCGTCAGGTG	CCCGATGCCGAGGTTGT	TGAAGTCGATGTCTTA		
18	CCAGCAAGCGATGGAG	CTTTTCCTACTTCGGCG	CTAAAGTTCTTCACCC	CCGCCACCATTACCCCC		
19	ATCGCCAGTTCAGAG	TCCCTTGCCCTGATTAA	AAATACCGGAAATCCT	CAAGCACCAAGGTACGC		
20	TCAATTGGTCCAGCCG	TGATGAAGACGAATTA	CCGGTCAAGGGCATTT	CCAACTGAATAACAT		
21	GGCAATGTTACAGCTT	TCTGTCGGGGATGA	AAGGATGGTCCGGCAT	GGCGGCGCGGCTCTTT		
22	GCAGCGATGTCACGCG	CCCGTATTTCGCTGGT	GCTGATTACGCAATCA	TCTTCCGAATACAGCA		
23	TCAGTTTCTGCGTTCC	ACAAAGCGACTGTGTG	CGAGCTGAACGGGCAA	TGCAGGAAGAGTTCTA		
24	CCTGGAACTGAAAGAA	GGCTTACTGGAGCCGC	TGGCAGTGACGGAAACG	GCTGGCCATTATCTCG		
25	GTGGTAGGTGATGGTA	TGCGCACCTTTCGCTGG	GATCTCGGCGAAATTC	TTTGCCGCACTGGCCC		
26	GCGCCAAATATCAACAT	TGTCGCCATTGCTCAG	GGATCTTCTGAACGCT	CAATCTCTGCTGTTGT		
27	AAATAACGATGATGCG	ACCACTGGCGTGGCGG	TTACTCATCAGATGCT	GTTCAATACCGATCAG		
28	GTTATCGAAGTGTTG	TGATTGGCGTGGGTGG	CGTTGGCGGTGGCGTG	CTGGAGCAACTGAAGC		
29	GTACGCAAAAGCTGGCT	GAAGAATAAACATATC	GACTTACGTGTCTGCG	GTGTTGCCAACTCGAA		
30	GGCTCTGCTCACCAAT	GTACATGGCCCTTAATC	TGGAAAACTGGCAGGA	AGAACTGGCGCAAGCC		
31	AAAGAGCCGTTTAATC	TCCGGCGCTTAATTCG	CCCTCGTGAAGAAATAT	CATCTGCTGAACCCGG		
32	TCAATTGTTGACTGCAC	TTCCAGCCAGGCAAGT	GCCGATCAATATGCCG	ACTTCTGCGCGAAGG		

$A_F(i, j) = \text{asc}(T_F(i, j))$

【図 10】

試料 F				C1F (i)	C2F (j)	C3F (j)
i	1	2	3	4		
1				11151500000617060406170017150200	eee6ae0de69f53cea8f58dd19aceb3	2c54d2518734f909fb5cfc45345a3f0a
2				0606020602150015150600400151300	3081ce3ca299962e5c3ca3230f4e0f1	f95d0609854fc84b8cc10e1a10d65390
3				04060615000015060206170002001515	7c69c0c1d8309ed2c8c0c9f39dca70e4	77351858c115a9a08189540c50f3985
4				0204001702020611511021513000200	26d6e260d6fef0ec1dac32d0dcd4c8c1	7e3ef48a38fa9b59ddc13d3f6a00820
5				1113150200060606001711502131100	28ae98c248d0c0c8ee887e6479627e3b	e0db502a78e8e2ff4d5274570432607
6				1704020211317110013060206021106	9cead396d6eecc5847d4d4ced2c8deb3	f3fa80314bd4748d873e603e6af9f4ca
7				1500170410606110011000213111306	fedefaced169e2c5defcc6e3ad3bc	59876afe034e68db3e6ec0f7b061cb
8				13130006170200151706040215021504	51d6f0c0f6d6ae2ae8efad63262dfc1c3	34d59818d48bead7db8a66174a6eb5e3
9				1517150000060613000400600171506	0eb7a499d4fcd27ad2cd2c8dbb18e92bc	2991c9cda1809f2aaa39a1c5dcd4ecdd
10				04020415130006130004000617020402	52ce75af7e0d28acaf23a3c18c2c2ff	5a076c115266a7defbae0a0c462a1b87
11				15021313001517170000041100110215	eaefcd68a0d0ddcc44ebfed1b2fcd6c69	283fe85987d2ecb336e0bef404388566
12				02131511170604041711000013040000	58fee68e9cd3eeb8d1eed2dcccdb8c	b8345d7588a0b6ff2f4feabf6ffa04c
13				13150600021506061115060613130000	58cd3c84dac1469d17a81d2d6cd65c7	0c990828ab56535f075d6bd4a82de6
14				11170017130402000602021315060402	fa9cfa72d5fe4684bd4cccd5d5d86dc9	5559bf3afcb3482c0163efcd3db3d06
15				00061313021302151504150017130200	4ed2a286d67c32f1191f8d288dcd6c7	c26acfaa72413f4f4dd3267d1f69e17
16	X1			0002150415170006171151700131502	4efef4f8183ae8b9ae5cc7d6270fa	c38f103236420e20d55f8cb378f3868
17	X2			15131313131700151300020406000413	164386867cf54ecde5c34f8d2c6fe6f	b160c9cf21f8146cbdfel9886a23bb66
18				00020002170013171302060411130000	5addedd6c5b21c6a9eac20cef0db6a0d3	86a8be9ebc00476b5784e80ed3d586a3
19				17021502171506020604151517001717	0cf2feb066ec3d6c0d8ecf4aacaddf7	2ba6f503438cc16ee8ba76d6de0alddc
20				06020006021706001513061306060400	b2b0a0cac47d2d2e562d229ebdd4aea	a3078606253a6a678846a21d2160baee
21				150604061302131317020606130215	a17d40c676f886fdd5aae3e8e8ecfed1	b4832e73fb9cbefdddec0f9f8b78fdd
22				13041317040204110006111513020017	64daa2e4d0c26de8f3828ef47634d27f	6abace56982f3d724185d6433a18af27
23				040213131706130411001100020002	464ccd4376ccc6e3c2c5ble1d2dce278	2716a8643fb62f951a7ddc556681ef7c
24				02061100150617061115041517040017	64d866d870d4667dfc8cc4d0d86dca7	12bd48d38557b4d37d8e1c3f730fac2
25				0604040611131702000400004040600	5c91a4fbed6e8fbc6e6d499fe37d51	81dc839f9858b261efeealccba1b487b
26				13040002040613060215020400130217	ad4e246f8d24382798c34da36e0dae9	aaalef35889f618174dc18bela527abd
27				130204170613110215111715170213	26c2ce21c0c8c5d0926aee9cc16034a9	f62afad019b28da94b5a7558b04c18d9
28				040002150404040004000406060013	4ab1fbcdf6da4d8d0deae8b2b65f9	cd6d616cb3c827868bffa2184c02dfb
29				11150604021700150415040606060000	41ece2a4798ed298c4ffe363cf671ef	4577e7877da6750e0f5861209a754f7
30				15110002151100000006150017041315	49eee4ceceeee299fa1f5bdeeddcb893	ab55e00276b5ebccbb9b01821d63b347
31				17110217171300151500171500060015	22eda2f6c9cf3d062caaa70de3cc0ff	f39c84d4957540d74afadb38ba0828fd
32				04110600040615041100171704111706	6edecad2b9d2a4daed7d82c96d78fcea	01f43a66cd9ac887987c9ce903828f629
B1F (i)	1	1	1	170015020015151710411041715131513	0606171700001102	
B2F (i)	1	1	1	131304000202130410402131504041511	1100041317040000	
B3F (i)	1	1	1	b9552e59d8b8e8a8159d9c9ced8917afd17b9c721cf2814474	c3dde65557dd039c	
	1	1	1	cd849e54becac8e1825d711630ac64701e3fc85c378fec0e	e2f606c26e5224ed	
	1	1	1	ce9dc74c5191bef4103959294b8c2898f1f6e21ff9a71f23	5ab7a51c44d0e9b	
	1	1	1	82c9fe3e0f1e17b81c659b64602915791154f20aa364036e6c	032628431a02303f	

【図 11】

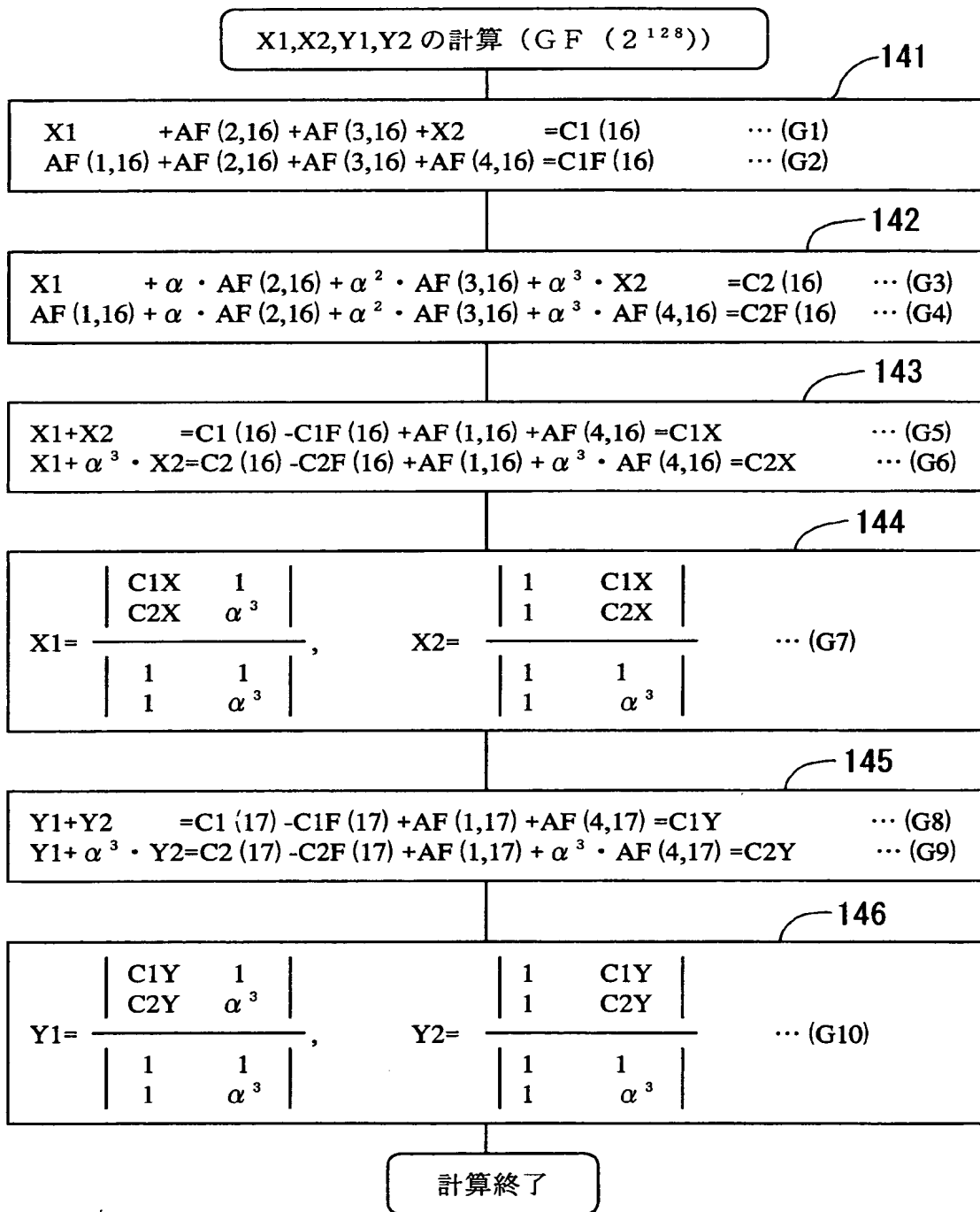
標準試料 E と試料 F との相違部					C1F (i)	C2F (i)	C3F (i)
j	1	2	3	4			
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							
30							
31							
32							
B1F (i)							
B2F (i)							
B3F (i)							

X1=hex (4347544747434754  
 4347544354434154)  
 → chr (X1) =TACTCTGCTGCGTGC  
 =T (1, 16) =TF (1, 16)  
 X2=hex (4754544743414747  
 4341474754545441)  
 → chr (X2) =ATTGGACGGCGTTG  
 =T (4, 16)  
 Y1=hex (4754434341544154  
 4354474747474341)  
 → chr (Y1) =ACGGGCTCTATACCTG  
 =T (1, 17)  
 Y2=hex (4154434354475441  
 4743544741414754)  
 → chr (Y2) =TGAAGTCGATGTCCTA  
 =T (4, 17) =TF (4, 17)

B1F (i)  
 B2F (i)  
 B3F (i)



【図 12】



【図 1 3】

標準試料 E									
j	i	1	2	3	B ( i , j )		4	5	C1B ( i ) ↓ C3B ( j )
1	1	1bfe3ed2d82560cd	eeddd4f00011ded3	186fb42d7cada747	03c03fcf4bc5e2c0			cbf0a0cc58c66212	
2	1	13003c8472208e8d	0263c628a3ca28a3	8a3ca217097f556d2	672140884001a98a			d21d95bffff192805	
3	2	c245c20a36477d07	e5972387580d8427	fed9df6933ed4060	da16156175a29ebb			b6aa68038a0a2b5d	
4	4	64d3d000a3c65a14	dbf2a0ce1936909c	ffda42ff4952e69a	6a1a55fa9b5983d0			0bf9e4e143f6a030	
5	5	08deb6358f1fd55	876a5318e09b66d3	f69d644037939a3c	d69673c419997882			7cb7ce4e979002db	
6	6	587558f2b90eca79	33db47a2a69cf658	1a63e96d388d76d3	585fe29a5c340059			0b5d76f526097e92	
7	7	cb6bd976d6dadfc9	9a4f7d913f52527d	255eccad92a6785d	a93645fd07937ac			a1419351bfacbe59	
8	8	b01fbc2aa628f2aa	39a87e84eaf6b4f0	032940eb818a1726	e3d78869d341243c			a5e0563fa0e0c23	
9	9	5837e19fed7a5534	054d7963596667bf	61937899a9cfe9d7	6d3c9838efa43218			elfed9fa20192ddd	
10	10	91b42560d85047ec	ad42d0105bcb51a6	d61d25096d68f3b9	75c5d35cd98af675			4ee5903efda62d6a	
11	11	66833823de68f6e1	53bed09b83bb79d7	030934d928b59d99	f2e384db7e0ca4e1			7ce41dfdd3d67975	
12	12	9f59766db5182d06	78601b5b410c08ce	4bc9ded977da0b90	5bb6e2837235af0e			d402d61410b5981a	
13	13	011a7f0ee566f0f9	ae7404338edb42a5	e3df4b62fa1a161d	653833692fad9905			0000000000000000	
B1B (i)									
B2B (i)									
B3B (i)									

【図 1 4】

		試料 G				TG (i, j)	
		1	2	3	4		
j \ i							
1		MRVLKFGG	TSVANAER	FLRVADIL	ESNARQQQ		
2		VATVLSAP	AKITNHLV	AMIEKTIS	GQDALPNI		
3		SDAERIFA	ELLTGLAA	AQPGFPLA	QLKTFVDQ		
4		EFAQIKHV	LHGISLLG	QCPDSINA	ALICRGEK		
5		MSIAIMAG	VLEARGHN	VTVIDPVE	KLLAVGHY		
6		LESTVDIA	ESTRRIAA	SRIPADHM	VLMA GFTA		
7		GNEKGELV	VLGRNGSD	YSAAVLAA	CLRADCCE		
8		IWTDVDGV	YTCDPRQV	PDARLLKS	MSYQEAME		
9		LSYFGAKV	LHPRTITP	IAQFQIPC	LIKNTGNP		
10		QAPGTLIG	ASRDEDEL	PVKGISNL	NNMAMFSV		
11		SGPGMKGM	VGMAARVF	<b>AAMSRARI</b>	SVVLITQS		
12		SSEYSISF	CV PQSDCV	RAERAMQE	EFYLELKE		
13		GLEPLAV	TERLAIIS	VVGDMRT	LRGISAKF		
14		FAALARAN	INIVAIAQ	GSSERSIS	VVVNNDDA		
15		TTGVRVTH	QMLFNTDQ	VIEVFVIG	VGGVGGAL		
16		LEQLKRQQ	SWLKNKHI	DLRVCGVA	NSKALLTN		
17		VHGLNLEN	WQEELAQA	KEPFNLGR	LIRLVKEY		
18		HLLNPVIV	DCTSSQAV	ADQYADFL	REGFHVVT		
19		PNKKANTS	SMDYYHQL	RYAAEKS R	RKFLYDTN		
20		VGAGLPVI	ENLQNLLN	AGDELMKF	SGILSGSL		
21		SYIFGKLD	EGMSFSEA	TTLAREMG	YTEPDPRD		
22		DLSGMDVA	RKLLILAR	ETGRELEL	ADIEIEPV		
23		LPAEFNAE	GDVA AFMA	NLSQLDDL	FAARVAKA		
24		RDEGKVL R	YVGNIDED	GVC RVKIA	EVDGNDPL		
25		FKVKNGEN	ALAFYSHY	YQPLPLVL	RGYGAGND		
26		VTAAGVFA	DLLRTL SW	KLGV0000	00000000		

(0=hex(00))

AG (3, 11) = asc(TG (3, 11))  
 = asc(AAMSRARI) = hex(49524152534d4141)

【図15】

j	試料G				C1G(j)	C2G(j)	C3G(j)
	i	1	2	3	4		
1					080c12161alc1ela	b00aecf04c2ef99f	853157c738293132
2					1c0a1f0506101611	4060bd7f361cd9be	53e62c6d43e1e330
3					100f031502161506	ba079d25ec2e0265	018e9a852168de0a
4					1b0f091b1f1f0119	311b1d8f0d2d56c8	5acal19f94e4fc75
5					15171d0908160706	c12d7bfb220e0493	9ae7b9a9da809bfd
6					0c140f021703080c	76ec2b0a7f4b2d0d	d89182e718758c0b
7					161d0db19111d0b	6e2f236535e502d2	82ad3d0300e7f125
8					16101b0f030f140d	025a032baf91ddb	2950d69f30b2a756
9					150106161c131305	9b3f18a076177af2	37180b557f0dca49
10					11111d1505040a0e	bfd37f7f1d48239b	8c49d625224a62cf
11					11120c1719061717	899af6211deeda4e	316a57d5c94eb8ff
12					100a0c0416090207	5c287a1236a10dfc	8f592d4e95bca35
13					17110905041e0d09	072b23eb700ea0b2	b059bcd9f19e8e52
14					0d0d0c1c110d0ale	2f1168007f959d57	bcd95a7354b4a3e8
15					1218131d10091705	26384b31822d0038	06dc9efcc1d40600
16					171b120a10040d15	178952505234e2f4	ae58ade31b5d32d8
17					04160a1a03001506	921c58fc5bb45ecf	0534866193eee904
18					1818150a020e0e1f	9ec4e15a664a2d6c	44d0bbe636f35d6a
19					030209041f081103	45b60fbe63220e12	5e77de978ae3f19f
20					0d02161d1f000901	03c612c16374207e	702f4f4ae995f420
21					06160d17040dle1b	62d6dd69c4139b5c	b72f416a09e849e1
22					090201081c111712	9bf61d7278696aaf	ec07e7981ab535c3
23					09030d1d07051903	77691de9bf692c92	cb3b1575ea29819c
24					1b101d1alc051209	1ddc0306603da192	3ec5ff16b96c8f80
25					1f151f06061e110c	670b217812c06415	791c2e2764be846a
26					16151a13454a5459	b8b382bab8cfdfd	8b582b47680fa178
B1G(i)	0202080a1f061707	0415000b06010f0e	47454040a1d0d01	5b444d59555c5a41			
B2G(i)	009ble6f9ce9cef5	5de8aee40ea7c7a8	459da64a0de38bc3	8929a2526c681df5			
B3G(i)	bd5e9e80ed61ea3e	b5494226607bdccc	55358c5ae2ccb4be	5d7c5f564db1f1da			
	i	1	2	3	4		

【書類名】 要約書

【要約】

【課題】 ヌクレオチド又はアミノ酸などの生物学的物質の配列情報を少ないデータ量で近似的に記録する。

【解決手段】 標準試料EのDNAを構成する一列のヌクレオチドの配列を示すテキストデータを所定の変換規則に従ってバイナリーデータに変換し、このバイナリーデータを複数行で複数列の $m$ ビットの部分データ $A(i, j)$ に分割する。各行の部分データ $A(i, j)$ を非配列方向にガロア体 $GF(2^m)$ 上で演算してパリティ $B1(i) \sim B3(i)$ を求め、各列の部分データ $A(i, j)$ を配列方向にガロア体 $GF(2^m)$ 上で演算してパリティ $C1(j) \sim C3(j)$ を求め、これらのパリティ情報によってヌクレオチドの配列を近似的に表す。

【選択図】 図8

出 願 人 履 歴 情 報

識別番号 [300000513]

1. 変更年月日	2001年 7月13日
[変更理由]	住所変更
住 所	埼玉県さいたま市西堀4丁目11番7号627
氏 名	大森 聡